DB 20273 BE-5180-(2)

B9

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number: **0 679 716 A1**

(12) **EUROPEAN PATENT APPLICATION published in accordance with Art. 158(3) EPC**

(21) Application number: **95900295.7**

(51) Int. Cl.[6]: **C12N 15/11**, C12Q 1/68, //G01N33/566

(22) Date of filing: **11.11.94**

(86) International application number: **PCT/JP94/01916**

(87) International publication number: **WO 95/14772 (01.06.95 95/23)**

(30) Priority: **12.11.93 JP 355504/93**

(43) Date of publication of application: **02.11.95 Bulletin 95/44**

(84) Designated Contracting States: **AT BE CH DE DK ES FR GB GR IE IT LI LU MC NL PT SE**

(71) Applicant: **Matsubara, Kenichi**
Room 804, 18-1, Yamadahi-gashi 3-chome
Suita-shi,
Osaka 565 (JP)
Applicant: **Okubo, Kousaku**
11-26, Segawa 2-chome
Minoo-shi,
Osaka 562 (JP)

(72) Inventor: **Matsubara, Kenichi**
Room 804, 18-1, Yamadahi-gashi 3-chome
Suita-shi,
Osaka 565 (JP)
Inventor: **Okubo, Kousaku**
11-26, Segawa 2-chome
Minoo-shi,
Osaka 562 (JP)

(74) Representative: **Vossius, Tilman et al**
Dr. Volker Vossius,
Patent- und Rechtsanwaltskanzlei,
Holbeinstrasse 5
D-81679 München (DE)

(54) GENE SIGNATURE.

(57) A 3'-directed cDNA library which accurately reflects the abundance ration of mRNA in a cell has been prepared from various human tissues, and sequencing of the cDNAs contained in the library has be conducted to examine the incidence of each cDNA in each tissue. As each cDNA has expression information with each tissue corresponding to the mRNA concentration, these cDNAs are usable as a probe or primer for detecting cell anomoly or discriminating cells. The cloned gene can produce porteins utilizable as a medicine or the like.

EP 0 679 716 A1

Fields of the Invention

The present invention relates to purified single-stranded DNA molecules, purified single-stranded DNA molecules complementary thereto or purified double-stranded DNA molecules consisting of said single-stranded DNA molecules, which can specifically hybridize to human genomic DNA, human cDNA or human mRNA at particular sites. The DNA molecules of the present invention can be used for detecting the overall or individual expression status of mRNAs coding for the corresponding cellular proteins, detecting and diagnosing cellular abnormalities due to disease and viral infection, or distinguishing and identifying the cell type, and efficiently cloning genes expressed in a tissue-specific manner. The present invention further includes cloned DNA molecules which can be used to produce proteins useful as pharmaceutical products or the like.

Related Arts

Recognizing the importance of the most fundamental attribute of mRNA, that is, "the nature of the cell is determined by the expression pattern of genes as reflected in the population of mRNA", the inventors of the present invention have proposed "body mapping" as a unique approach to their objective. This is an entirely novel attempt to prepare "the information on gene expression" for presumably about 200 different kinds of cells and tissues present in the human body and elucidate when, where and to what extent a certain gene is expressed, and map genes to the respective organ or cell type in which they are expressed.

While a variety of cells in the living body express various proteins depending on their respective biological functions, the intracellular concentrations of these proteins vary according to the cell type, stage of development and differentiation, environment, etc.

In general, genes are classified into "genes encoding proteins essential for the life of the cell" and "genes encoding proteins responsible for functions specific to the cell". Of these two, "genes encoding proteins essential for the life of the cell" are expressed constantly in all types of cells and also called "housekeeping genes", while "genes encoding proteins responsible for functions specific to the cell" are often expressed specifically in a particular type of cells or a particular group of cells, and also may be specifically expressed at a particular stage of cellular development and differentiation. Furthermore, they are often "inducible genes" and the amount of their expression varies depending upon the environment to which cells are exposed. In other words, cells may grow as a result of the expression of "genes encoding proteins essential for the life of the cell" and display their specific functions as a result of the expression of "genes encoding proteins responsible for functions specific to the cell".

However, under abnormal cellular conditions due to disease or infection, the expression of genes within individual cells is altered as compared with that under the normal conditions. Especially, during viral infection, RNAs encoding virus-specific proteins are synthesized in large amounts within the cell, leading to the production of said protein in large amounts. In other words, the alteration in the expression level of genes within the cell, especially as reflected in the concentration of intracellular mRNA, can lead to such abnormal cellular conditions as seen in diseases.

Thus, the function of each cell in the living body is closely related to the expression status of genes within the cell. Accordingly, in order to elucidate the function of each cell at molecular level or to investigate the pathogenesis of a disease at molecular level, it becomes necessary to comprehend the expression status of cellular genes, especially the intracellular concentration of each mRNA.

A theoretically possible approach to this objective is the extraction and analysis of all cellular proteins for determination of expression status. However, although it may be possible to isolate a specific protein, in most cases it is almost impossible to completely isolate all of these proteins, because a great variety of proteins are expressed within the cell.

Another approach is to directly estimate the concentrations of cellular mRNAs corresponding to all intracellular proteins. However, although it may be possible to isolate a specific mRNA, it is practically impossible to completely isolate all of these mRNAs and directly estimate their amounts, because a great variety of mRNAs are synthesized simultaneously within the cell and furthermore they may be unstable and susceptible to enzymatic degradation during their extraction.

This invention aims to provide DNA molecules which can be used as probes or primers required for detecting the overall or individual expression status of mRNAs coding for the corresponding cellular proteins, detecting or diagnosing cellular abnormalities due to disease or virus infection, recognizing and identifying various cell types, and efficiently cloning genes expressed in a tissue-specific manner. Moreover, the present invention aims to provide cloned DNA molecules which can be used to produce proteins useful as pharmaceutical products.

Summary of the invention

In general, the genetic information flows in order from DNA to mRNA and to protein (F. H. C. Crick, 1958). That is, "the information for the amino acid sequence of a protein" is first transcribed into mRNA and then translated into protein.

To explain this in further detail mammalian genes commonly comprise a region encoding a protein and a region regulating the expression of said gene. The regions of a gene encoding protein (called "exons") are often separated by intervening sequences (called "introns"). When a gene is transcribed into RNA, the introns of the precursor RNA (pre-mRNA) are excised and exons are connected in tandem to form a contiguous structure coding for a particular protein (this process is called "splicing"). On the other hand, the region regulating the expression of gene comprises, in addition to the regions directly regulating transcription such as a promoter and operator which are present upstream of the transcription region, untranslated regions are located both upstream (5') and downstream (3') of the coding region. In particular, 3' untranslated region (3' UTR) is important for regulating expression, since it contributes to the transport and stability of mRNA. During the processing of pre-mRNA, a methylated cap is added at its 5' end, the 3' untranslated region is cleaved at a specific site, a poly(A) tail is attached by adding 100 - 200 adenylate residues to the cleaved end, and the coding regions are spliced together to form mRNA. The protein is then synthesized after attachment of ribosomes to the mRNA.

The inventors of the present invention have elucidated that, in general, when the intracellular level of a particular mRNA is high, the expressed amount of the corresponding protein is also elevated, and also that it is possible to estimate the relative concentration of each intracellular protein by estimating relative intracellular concentration of the corresponding mRNA [DNA sequence 2, 137-144 (1991); Nature genetics, 2, 173-179 (1992)].

Basically in the present invention, mRNA is extracted from a particular cell and cDNA is synthesized by conventional methods using reverse transcriptase. However, in the present invention, cDNA is synthesized using a method developed by the inventors of the present invention so as to reflect the relative intracellular concentration of mRNA. A cDNA library is constructed and a group of cDNAs representing the population of total mRNA are cloned and sequenced.

An approach which appears to be similar to the one used by the inventors of the present invention but is entirely different, is the method of cloning of a cDNA library constructed by the random priming by Venter et al.

Venter's group randomly cloned cDNAs from commercially available cDNA libraries derived from brain cells (catalog No. 936206, 936205 or 935, Stratagene, California) and determined their base sequences [Science 252, 1651-1656 (1991); Nature 355, 632-634 (1992)].

While the method used by Venter et al. involves sequencing of cDNAs obtained by random priming, this method has the following drawbacks:

1) Since random cloning of various regions of a single-stranded mRNA may often lead to the formation of many cDNA fragments without any mutual overlapping portions, it is difficult to determine whether these cDNA fragments are derived from the same mRNA or a different one,

2) The longer a mRNA strand, the higher the chance for said mRNA to be reverse-transcribed into cDNA, and

3) Since the availability of each primer to be used among random primers differs depending on their base sequences, the relative frequency of cDNA synthesis is variable.

From aforementioned reasons, the relative frequency of appearance of cDNA does not reflect the relative concentration of cellular mRNA. Consequently, it is impossible to determine the relative concentration of each mRNA and the actual population of intracellular proteins by using the method of Venter et al.

However, with the method developed by the inventor of the present invention, it is possible to construct a cDNA library which precisely reflects the relative concentration of mRNA without any of the aforementioned complications. Since, in the present invention, cDNA is synthesized using only "poly-T" as the primer, the 3' ends of the cDNA have "a poly A tail". Therefore, the synthesis of cDNA with "poly-T" as the sole primer is initiated from the 3' end resulting in the formation of 3'-oriented cDNA. Since the 3' untranslated sequence is unique to a particular mRNA species and not present in other mRNA species [Birnstiel, M. L., et al., Cell 41, 349-359 (1985)], almost all the 3' end-oriented cDNAs hybridize with specific mRNAs. Digestion of the resulting cDNA with a restriction enzyme MboI which recognizes the specific four-base sequence GATC results in the formation of cDNA extending from the 3'-terminus to the first MboI restriction site. In the present invention, each cDNA thus cloned and included in "a cDNA library faithfully reflecting the relative intracellular concentration of mRNA" is called a "gene signature" (abbreviated as GS hereinafter). A GS includes not only the double-stranded DNA but also each single-stranded DNA thereof.

The present invention relates to a purified single-stranded DNA, purified single-stranded DNA complementary thereto, or a purified double-stranded DNA consisting of said single strands, containing all or a portion of a single-stranded DNA (or a single-stranded DNA complementary thereto) comprising any of the base sequences listed under the sequence identification number (SEQ ID NO) 1 - 7837 and hybridizing specifically to a particular site of human genomic DNA, human cDNA or human mRNA. The present invention also relates to probes and primers consisting of said single-stranded DNA. The present invention also relates to a purified single-stranded DNA, a purified single-stranded DNA complementary thereto, or a purified double-stranded DNA consisting of said single strands, containing all or a portion of a single-stranded DNA (or a single-stranded DNA complementary thereto) which is complementary to a human mRNA containing any of the base sequences listed under SEQ ID NO 1 - 7837 (wherein T is read as U) or any portion thereof at its 3' region and hybridizing specifically to a particular site of human genomic DNA, human cDNA or human mRNA. The present invention also relates to probes and primers consisting of said single-stranded DNA.

The present invention is explained further in detail as follows.

The DNA of the present invention not only includes a single-stranded DNA (or a single-stranded DNA complementary thereto) comprising any of the base sequences listed under SEQ ID NO 1 - 7837 but also includes a single-stranded DNA containing a portion of said single-stranded DNA (or said single-stranded DNA complementary thereto) if it hybridizes to human genomic DNA, human cDNA or human mRNA.

Furthermore, the DNA of the present invention not only includes a single-stranded DNA (or a single-stranded DNA complementary thereto) which is complementary to a mRNA containing any of the base sequences listed under SEQ ID NO 1- 7837 (wherein T is read as U) or any portion thereof at its 3' region but also includes a single-stranded DNA (or a single-stranded DNA complementary thereto) containing a portion of said single-stranded DNA (or said single-stranded DNA complementary thereto) if it hybridizes to human genomic DNA, human cDNA or human mRNA.

In addition, the DNA of the present invention not only includes a single-stranded DNA or a single-stranded DNA complementary thereto but also includes a double-stranded DNA consisting of said single strands.

Obviously, the term "contain" as used herein does not necessarily mean that the DNA of the present invention contains at a single site without interruption (1) "a single-stranded DNA (or a single-stranded DNA complementary thereto) comprising any of the base sequences listed under SEQ ID NO 1-7837 or a portion thereof" or (2) "a single-stranded DNA (or a single-stranded DNA complementary thereto) which is complementary to a mRNA containing any or any portion of the base sequences listed under SEQ ID NO 1 - 7837 (wherein T is read as U) at its 3' region or a portion of said single-stranded DNA." In other words, the term "contain" is applicable also to the case where one or more exogenous bases are inserted in the base sequence of the DNA (1) or (2).

The hybridization to a particular site of human genomic DNA, human cDNA or human mRNA can be achieved under standard conditions (see e.g., ,Molecular Cloning: A Laboratory Manual, Sambrook, J., et al., Cold Spring Harbor Laboratory Press, 1989). In the following preferred embodiment, there will be described methods for constructing a cDNA library which reflects precisely the relative intracellular concentration of mRNA, cloning cDNA groups which correspond to total mRNA, and determining the base sequence of each cDNA.

First, cells from specific tissues, for example, cells from organs, for example, cells derived from human liver (HepG2) are grown, and the total mRNA is extracted by standard procedures. mRNA thus obtained is attached to a vector to construct a cDNA library.

For example, mRNA is attached to the vector plasmid pUC19, which has the M13 sequences flanking the cloning site, as follows.

pUC19 is cleaved by HincII and PstI and poly-T of 20 bp - 30 bp is added to the PstI-digested end to which the 3'-end poly-A tail of the mRNA is hybridized (Fig. 1a). After the DNA strand is extended with conventional methods using reverse transcriptase, a double stranded DNA is formed with DNA polymerase (Fig. 1b). The double stranded DNA thus obtained is cleaved with the restriction enzyme MboI which recognizes a specific four base sequence (Fig. 1c).

MboI, which recognizes a four base sequence (GATC), cleaves the DNA within a few hundred bases from the poly-A tail. Since MboI is found to digest, without exception, about 300 human cDNAs which were randomly selected from the GenBank data base by the inventor of the present invention, this enzyme cleaves the cDNA to be cloned at a specific site. In addition, as pUC19 is prepared in dam$^+$ E. coli, e.g., E. coli JM109 and since its adenine at the MboI recognition site is methylated (G^{m}ATC), it is not cleaved by MboI.

Subsequently, in order to prepare a vector containing the double-stranded DNA which has previously been attached to pUC19 and has the MboI-cleaved end, the pUC19 DNA is digested with BamHI to make termini cohesive with the MboI-cleaved end. Since the recognition sequence of BamHI (GGATCC) contains that of MboI (GATC), the extended portion of the double-stranded DNA is not cleaved with BamHI.

The resulting double-stranded DNA is then circularized by standard ligation methods, and the recombinant vector plasmid thus prepared is introduced into E. coli, e.g., E. coli DH5 in order to make a cDNA library.

With this method, only a clone containing the base sequence upstream of the poly-A tail of the mRNA is obtained.

Since the average size of the inserted cDNA fragment is relatively small, 270 bp, it is free from biased cloning resulting from variations in the efficiency of cDNA synthesis and transformation that occur in the case of larger sized DNAs. Furthermore, because instability due to repeated base sequences and the like is eliminated, the cDNA library of the present invention faithfully represents the relative concentration of mRNA in the cell.

Furthermore, when the cDNA inserted into the vector is relatively short, it is possible to accurately amplify the cDNA fragment using the sequence of the vector flanking it as a primer. It is also possible to determine the base sequence from the 5' end directly by the PCR without interference from the 3' poly-A tail which will reduce the accuracy of sequence determination.

Amplification of the GS, i.e., the cDNA fragment inserted into the vector, is performed as follows.

The E. coli cells in which the cDNA library is introduced are grown using standard methods and lysed. Debris contained in the bacterial lysate are removed by centrifugation and the supernatant containing the vector DNA is recovered. The vector DNA thus obtained is used as the DNA template for amplification by the PCR (Fig. 1d, amplification with PCR primers 1 and 2).

Base sequences flanking both ends of the GS is properly selected for use as primers and the PCR is performed under standard conditions. PCR products thus obtained are subjected to the elongation reaction using fluorescence primers complementary to the vector sequence flanking the 5' end of the GS, and the sequence is determined with an autosequencer (Fig. 1d, sequence determination with dye primer).

Based on the results of the sequence determination of each GS, the species and the frequency of appearance of the GS in each tissue or cell type are analyzed.

As to each cell type not only normal cells but also cells under pathogenic conditions (such as tumor cells, virus infected cells, etc.) can be used without any restriction. For example, liver cells (from fetus, neonate or adult), various hematopoietic cells (granulocytic, monocytic, etc.), lung cells, adipocytes, endothelial cells, osteoblasts, colon mucosa cells, retinal cells and hepatoma cells (HepG2, etc.), and promyelocytic leukemia cells (HL60, etc.) will be used. The appearance frequency for each GS is described for each cell type in Tables 1 through 219. There, patent number represents "SEQ ID NO for each GS", size represents the "length of each GS", and F represents the "sum of appearance frequencies in the cells studied". In addition, hepG2 stands for "hepG2 (a liver cancer cell line)", HL60 stands for "HL60 promyelocytic leukemia cell line", granulo stands for "granulocytoid, HL60 stimulated by DMSO", mono stands for "monocytoids, HL60 stimulated by TPA", 40 w liver stands for "40 w neonatal liver", 19 w liver stands for "liver of a 19 weeks old fetus, adult liver is "adult liver ", lung stands for "adult lung", adipose stands for "subcutaneous adipose tissue", endothel stands for "primary cultured aortic endothelium", osteoblast stands for "primary cultured osteoblast", colon mucosa is "colon mucosa", small cell carci stands for "small cell carcinoma of lung", retina is "retina", cerebral cortex is "cerebral cortex", adenocarci (lung) stands for "adenocarcinoma of lung", squamous cell ca (lung) stands for "squamous cell carcinoma of lung", keratinocyte stands for "primary cultured keratinocyte", fibroblast stands for "primary cultured fibroblast", Alzheimer stands for "Alzheimer temporal lobe", cerebellum stands for "cerebellum", visceral fat is "visceral fat", corneal epithelium is "corneal epithelium", peripheral granulocyte is "peripheral granulocyte", neuroblastoma is "neuroblastoma" and taste bud of tongue is "taste bud of tongue".

"Accession number of target mRNA" represents the accession number of the entry in GenBank Release 79 whose base sequence has homology with that of each GS, "match %" represents the percent homology of the GS sequence relative to that of said homologous sequence, "match starts at (GS)" represents the base position counted from the 5'-end of the GS at which the region for homology calculation starts, "match starts at (GenBank)" represents the base position counted from the 5'-end of the GenBank sequence at which the region for homology calculation starts; and "GenBank target size" represents the whole length of the GenBank sequence corresponding to the GS. The columns in Tables 1 - 219 represent the same items as in Table 1.

Based on the data in Tables 1 - 219, each GS can be classified into several groups. A GS, which is expressed at high frequency in a specific cell or groups of cells with similar property, for example,

promyelocytic leukemia cell, granulocyte and monocyte and not expressed entirely or expressed very little in other cells (groups), is a likely GS corresponding to the gene encoding "the protein responsible for functions specific to the cell" (e.g., GS0001553, GS0002047, GS004895, etc.). On the other hand, a GS, which is expressed commonly in every kind of cell, most likely corresponds to the gene encoding "the protein essential for the life of the cell" (e.g., GS0000019, GS0000155, GS000861, etc.). In addition, some GSs are expressed at low frequency (e.g., GS0000013, GS0002399, GS0003155, etc.).

Since the GS with the sequence determined as described above will reflect the population of mRNA expressed in a particular cell, it must be possible to find the relative concentration of mRNA in each cell by determining the appearance frequency for each GS in a cDNA library derived from that cell. Therefore, to confirm the correlation between the appearance frequency for each GS in a cDNA library and the relative concentration of cellular mRNA, the GS thus obtained was labeled with ^{32}P by standard methods and used as the probe in the following hybridization test. mRNA isolated from a specific cell is hybridized to said ^{32}P-labeled probe under standard conditions. The results of this Northern hybridization test were such that, when a GS found with high appearance frequency in a cDNA library was used as a probe, a dense band was formed, confirming the correlation of the frequency of appearance of the GS with the relative concentration of mRNA in the cell (see Example 5).

Similarly, the colony hybridization test of the cDNA library constructed as described above with a ^{32}P-labeled probe prepared as described above showed a close correlation between the frequency of appearance of the GS and the number of colonies hybridized with said GS (see Example 6), confirming the correspondence of the frequency of appearance of the GS and relative concentration of the GS in a cDNA library.

From the above results, by determining the appearance frequency of each GS in a cDNA library derived from a variety of cells, it has become possible to determine the expression status of the gene (or mRNA) corresponding to each GS. This fact implies conversely that each GS may be useful for industrial purposes as a specific probe or primer encoding information about the expression status of its corresponding gene (or mRNA) for each cell. For example, when it is proven that "a certain GS appears at high frequency only in a cDNA library derived from tissue A, that is, the gene corresponding to said GS is specifically expressed only in tissue A", by conventional cloning of the corresponding full-length cDNA using said GS as a probe or primer, it is possible to clone a full-length gene which is expressed in a tissue-specific manner.

Furthermore, for example, when it is proven that "the frequency of appearance of a certain GS is low in a cDNA library derived from tissue B, that is, the appearance frequency of the gene corresponding to said GS is low in tissue B", by examining the expression frequency of the gene corresponding to said GS in a test sample of tissue B from a patient using said GS as a probe or primer, it may be possible to identify the pathogenic gene, wherein an unusually high expression frequency of said gene being a strong indication that said GS may be the gene involved in the pathogenesis. Furthermore, by conventional methods for cloning said full-length cDNA using said GS as a probe or primer, it is possible to isolate said pathogenic gene and elucidate its characteristics.

In practice, the DNA of the present invention may be used as a probe or primer for detecting and diagnosing disease, cloning a pathogenic gene or related gene, cloning a viral gene, identifying and recognizing cell types, cloning a species-specific promoter and gene mapping.

One GS corresponds to one mRNA. It is therefore obvious that any portion of cDNA complementary to each mRNA carry the same "information for expression" as the GS. Accordingly, the DNA of the present invention is not restricted to "the DNA comprising the GS itself or portion thereof", but also includes the DNA comprising, for example, "a full-length cDNA complementary to each mRNA" and "the non-GS region of the cDNA complementary to each mRNA or a portion thereof". They can be used as a probe or primer comprising the same "expression information" as that of the GS and can be used as a probe or primer in a similar manner as a GS. For example, by using a GS or a portion thereof as a probe or primer, it is obviously possible for those skilled in the art to readily isolate "a full-length cDNA corresponding to each mRNA" or "the non-GS region of the cDNA complementary to each mRNA or a portion thereof". For example, as described hereinafter, conventional techniques such as "5' RACE", "nesting" and "inverse PCR" can be used.

An example of the method for detecting disease using the GS of the present invention will be described. As shown in Tables 1 - 219, with the method described above it is possible to detect a GS present specifically in a cDNA library constructed from each tissue by detecting and comparing the frequency of appearance of GS in each tissue. It is also possible to identify a GS corresponding to a protein which is expressed commonly in various tissues or which is expressed at low frequency. These GSs are denatured and then fixed on an appropriate filter, for example, nylon filter or nitrocellulose filter. It is

convenient to use a single filter with many GSs fixed on it. Usage of a single filter on which many denatured DNAs are fixed is well known. An example may be "the Escherichia coli Gene Mapping Membrane" (Takarashuzo, code No. 9035). It is a single nylon filter on which the cosmid contigs of genomic DNA of E. coli are fixed. It is possible to prepare a filter comprising a group of specific GSs corresponding to proteins expressed in a particular tissue, a filter comprising a group of GSs corresponding to proteins commonly expressed in various tissues, or a filter comprising a group of GSs corresponding to proteins expressed at low frequency. The single-stranded GSs fixed on these filters are then hybridized to labeled complementary DNA fragments synthesized using "random primers" prepared from template mRNA extracted from a test tissue, using four labeled nucleotides and reverse transcriptase (labeled mRNA can also be hybridized to the filters). Similarly, labeled complementary fragments synthesized using mRNA extracted from normal tissue as the template are hybridized (labeled mRNA can also be hybridized to the filters). If the profile of hybridization to a group of GSs has been categorized beforehand by comparing the hybridization profile of various pathogenic tissues to that of corresponding normal tissues, it is possible to diagnose the pathogenic condition of a particular test tissue by comparing the hybridization profile of the test tissue with that of the corresponding normal tissue and assigning that profile to a certain category. Virus infection can be detected in the same manner as in the case of other diseases.

Next, an example of the method for cloning pathogenic genes or their related genes using the GS of the present invention is described. As described above, using the filter on which denatured GSs are fixed, the GS-hybridization profile of various pathogenic tissues and that of corresponding normal tissues are compared. A considerable difference in the hybridization intensity between normal and pathogenic tissues will be an indication that the particular GS corresponds to a pathogenic gene. If a filter comprising only GSs specific for a particular tissue is applied to a sample from that particular tissue, the probability for detecting the GS with a great difference in hybridization intensity is elevated. Also a filter comprising GSs corresponding to proteins whose expression is low will facilitate the identification of the GS corresponding to the pathogenic gene by detecting an intense signal, because the hybridization signal for these GSs is usually weak. Once a GS corresponding to a pathogenic gene is found, said pathogenic gene can be cloned by established methods such as genomic Southern hybridization using said GS as a probe and/or a primer.

Furthermore, a method for cloning a full-length gene using a GS as a probe or primer is described in detail. Cloned genes isolated in the present invention are also appropriate for use in the production of proteins useful as pharmaceutical products. mRNA is extracted from tissues by conventional methods and cDNA libraries are then prepared (See Molecular Cloning, 2nd ed. Vol. 2, Section 8 New York; Cold Spring Harbor Laboratory). In this case, it is desirable to extract mRNA from tissues in which the target gene is highly expressed. One method to detect a specific gene in libraries thus prepared is, for example, to select positive clones via hybridization using a whole or partial GS as a probe. In general, since a GS is specific for a particular mRNA, hybridization can be carried out under certain stringent conditions. Probes used are at least more than 25 bases long, preferably more than 50 bases long, and more preferably more than 100 bases long.

Furthermore, if cDNA libraries, in which the cDNA for a specific gene is concentrated, are prepared, they will be preferable for selecting said specific gene. One method useful for this purpose is carried out as follows: 1) preparation of an affinity chromatographic column of resin on which the denatured GS corresponding to the specific gene is fixed; 2) application of mRNA extracted from a tissue to said column and retention of the mRNA species corresponding to the specific gene on said column; 3) elution and concentration of said retained mRNA; and finally 4) preparation of cDNA libraries using said concentrated mRNA species as the template. Another method is the selective amplification of cDNA corresponding to the specific gene by the PCR. Selective amplification of a specific gene is carried out as follows: using a partial sequence of a GS localized toward the 3' end of the specific gene as primer, cDNA is synthesized from mRNA with reverse transcriptase and 4 NTPs. To the 3' end of a single-stranded cDNA thus obtained a homopolymer such as poly-T is attached by the action of "terminal deoxyribonucleotide transferase (TdT)". In addition, using "a primer complementary to the homopolymer" and "a primer used in said reverse transcriptase reaction, or a primer whose sequence is included in the same GS but is located proximal to the 5' end", cDNA corresponding to the specific gene may be selectively amplified by the PCR [see 5'RACE (5' Rapid Amplification of cDNA ends): PNAS, Vol. 85, pp. 8998 - 9002 (1988); Nucleic Acids Res., Vol. 17, pp. 2919-2932 (1989)]. In addition, instead of the attachment of a homopolymer, there is another method comprising the following steps: 1) a single stranded anchor DNA is linked to the 3' end of a single stranded cDNA using "T4 DNA ligase"; and 2) said cDNA is amplified by the PCR using a primer complementary to said anchor DNA [Nucleic Acids Res., Vol. 19, pp. 5227-5232 (1991)]. Said primer is desirably more than 13 bases long, preferably more than 15 bases long, and more preferably more than 18

bases long. Furthermore, in order to enhance the efficiency of heat denaturation in the cycling reaction, said primer is preferably less than 50 bases long and more preferably less than 30 bases long. By linking said amplified DNA to a vector, a cDNA library concentrated with respect to the target gene is prepared.

In addition, it may be also possible to isolate a cDNA clone corresponding to the specific gene directly from the PCR products. Specifically, the PCR products are first separated by gel electrophoresis, subjected to Southern blotting analysis using the denatured GS as a probe, and examined for the presence of a band which specifically hybridizes to said GS. If a GS-hybridized band is detected, it is highly possible to isolate the cDNA clone corresponding to the specific gene by excising said band from the gel and subjecting it to direct cloning.

As described above, in order to further amplify the specific gene previously amplified by the PCR, it may be possible to perform the second PCR of the primary PCR products by replacing either or both primers previously used with a primer having the base sequence internal to said two primers (nesting) (Journal of Virology, Vol. 64, p. 864 (1990)). Nesting may be performed directly upon the products of the primary PCR. Alternatively, if a band which specifically hybridizes to the GS is detected by the Southern blotting analysis of the primary PCR products, nesting may be performed for the DNA obtained by excision of the band followed by extraction. In the case where a band which specifically hybridizes to the GS is detected by the Southern blotting analysis of nested products using the denatured GS as a probe, it is highly possible to successfully isolate the cDNA clone corresponding to the target gene by excising said band from the gel and subjecting it to direct cloning.

The isolated cDNA clone corresponding to the target gene may often correspond to the full-length mRNA, but it may be a cDNA with the 5' end deleted. In the case where the 5' end is deleted it is possible to isolate the full-length cDNA clone by conventional methods. For example, by screening a cDNA library using a probe comprising the base sequence in the 5' end region of the cloned cDNA, since the target position of said probe is shifted further toward the 5' end of the full-length cDNA than in the case of using a GS as a probe, it is possible to isolate only longer cDNA clones as the positive clone. Also by synthesizing cDNA using "a primer comprising the base sequence in the 5' end region of the cloned cDNA" with mRNA as the template followed by PCR amplification of "a single stranded cDNA having a homopolymer or anchor DNA sequence at the 5' end" and using" the primer used for previous cDNA synthesis or a primer having the sequence internal to that of said primer" and "a homopolymer or a primer complementary to anchor primer" as described above for the 5' RACE method, only the sequence toward the 5' side of the cDNA may be selectively amplified since the position of said primer is shifted further toward the 5' side of the full-length cDNA. Even if the cDNA thus obtained has a deletion at the 5' end, the population of cDNA fragments covering the full-length of the long cDNA may be obtained by repeating this procedure. It may be easy for those skilled in the art to obtain a full-length cDNA by suitably linking said cDNA fragments having overlap segments together.

Alternatively, by performing the inverse PCR (Inverse PCR: Genetics, Vol. 120, p. 621 (1988); Molecular Cloning, 2nd ed., Vol. 2, 14.12-14.13 (New York; Cold Spring Harbor Laboratory)), it may be possible to isolate a cDNA clone extending externally from the GS, that is, in the genomic DNA region. Specifically, the target DNA (genomic DNA or cDNA) is digested with restriction enzymes into fragments of about 2-3 kb and then circularized by ligating the cleaved ends. By performing the PCR for said DNA using "a set of primers which are complementary to the cDNA clone isolated using the GS or the GS as a probe or primer, and thereby making the direction of DNA synthesis mutually opposite (outward), it may be possible to amplify the DNA region extending externally from the GS. There is known a method to isolate a full-length genomic DNA of a specific gene by repeating this procedure (Nucleic Acids Res., Vol. 16, p. 8186 (1988)).

In addition, although "Taq polymerase" is conventionally used in the PCR described above, the cloning procedure may be more efficiently performed using the "LAPCR (long and accurate PCR" technique (Nature Genet., Vol. 7, p. 350-351 (1994), Nature., Vol.369, p.684-685(1994)).

Furthermore, needless to say that by linking said full-length gene thus obtained to a suitable expression vector followed by its expression in an appropriate host, it is possible to obtain the corresponding gene product (Molecular Cloning, 2nd ed.).

Next, there will be described an example of the method for identifying and recognizing cell types using the GS of the present invention. As shown in Tables 1 - 219, based on the appearance frequency of GS in each tissue and its comparison among tissues, it is possible to identify those GSs specifically present in a cDNA library constructed for each tissue. These "tissue-specific GSs" are fixed on a filter. It will be more convenient if GSs specific to each tissue are collected and fixed on a filter as a whole (e.g., a GS block specific for hepatocytes or cerebral cortex cells). As described above, to this filter are hybridized labeled complementary fragments synthesized using "random primers" prepared from mRNA extracted from test tissues or cells, "nucleotide containing 4 labeled nucleotides", and "reverse transcriptase". (Directly labeled

mRNA can also be hybridized to the filters.) Depending on the type of tissues or cells, intense hybridization signals will be observed with the GS groups specific to said tissue or cell. Furthermore, a tissue-specific promoter can be cloned by structure analysis of the 5' upstream sequence through the cloning of the corresponding gene using established methods such as genomic Southern hybridization with the "tissue-specific GS" as the probe and/or primer.

These tissue-specific promoters thus obtained are useful for gene therapy in the future.

Gene therapy in a narrow sense aims to supplement the defective protein of patients using gene technology, and in this case it is necessary to express the exogenous gene in a desired tissue in a desired quantity. For this purpose, a promoter which is known to be expressed in a specific tissue in a desired quantity (in most cases a large quantity is desired) is highly useful. Although, at present, a virus promoter is often used, it can be inactivated by endogenous modification such as methylation. Promoters provided by tissue-specific GSs will be ideal substitutes for viral promoters.

There will be described the method for chromosomal assignment of DNA corresponding to the GS of the present invention using the probe derived from the GS obtained as described above.

First, the Southern blotting method will be described.

According to this method, for example, chromosomes are isolated from a lymphoblast cell line of human normal karyotype (e.g., GM0130b), and then a monochromosomal hybrid cell is prepared by introducing each human chromosome into non-human cells, such as rodent cells, and cultured on a large scale by standard methods. Then the DNAs extracted from said hybrid cells are digested with various restriction enzymes and subjected to agarose gel electrophoresis. Then, the electrophoresed DNAs are hybridized to ^{32}P-labeled GS prepared as described above and used as the probe. By identifying the hybrid cell the DNA of which is hybridized to said probe, it is possible to identify the chromosome in which the DNA corresponding to the GS of the present invention is present. Southern hybridization test of the total human genomic DNA using each labeled GS as a probe formed a single band corresponding to the GS, indicating that the DNA of the present invention can be used as a desirable probe for human genomic DNA. It is obvious that a desirable probe for human genomic DNA can be used also as a desirable probe for human cDNA and human mRNA.

A method similarly using the PCR to determine chromosomal localization of the GS of the present invention will be described.

To prepare most appropriate primers, base sequences are selected from the sequence of the GS in question by conventional methods, for example, by using the computer software OLIGO4.0 (National Biosciences) and the oligonucleotides (20-24mer) having the selected sequences are synthesized. The preferred size of the sequence to be amplified by the PCR is from 50mer to 100mer.

Using the primers thus synthesized and the chromosomal DNA extracted from the monochromosomal hybrid cell as such as the template, amplification by the PCR is performed in a conventional manner. Resulting PCR products are subjected to non-denatured acrylamide gel electrophoresis and stained with ethidium bromide for fluorescent detection. The sizes of these PCR products are then determined.

Chromosomal assignment is confirmed when the presence of a PCR product of correct size is confirmed.

It is evident that a chromosome or chromosomes in which the DNA corresponding to a GS is localized can be identified by using these procedures. It has also become evident that the DNA of the present invention can be used as desirable primers for human genomic DNA since a single band has resulted from amplification of the total human genomic DNA by the PCR using primers designed based on each tested GS. Obviously, a desirable primer for human genomic DNA is also a desirable primer for human cDNA and human mRNA.

Brief Description of Figures

Fig. 1 shows the preparation of 3' MboI cDNA library.

Fig. 2 shows the results of tests of primers. A shows the location of primers on the vector; and B shows the electrophoretic patterns of DNA fragments amplified using the primers (A). Primers used are as follows: lane 1, FW (-40)/RV (-14); lane 2, FW (-40)/RV (-36); lane 3, FW (-40)/RV (-71); lane 4: FW (-40)/RV (-29); and lane 5, FW (-47)/RV (-48). Artifacts are indicted by arrows.

Fig. 3 shows the electrophoretic pattern of PCR products using FW(-40) and RV(-14) as primers. The lane at the right end shows the electrophoretic pattern of size markers and the other lanes show the PCR products using FW (-40)/RV (-14) as primers.

Fig. 4 shows the mRNA concentration reflecting the frequency of appearance of each GS in the cDNA library: especially, Fig.s 4A - 4D; experimental results; Fig. 4E, photographs of colonies; and Fig. 4F,

summary.

Fig. 5 shows the appearance frequencies for various cDNAs in the 3'-directed HepG2 cDNA library.

Fig. 6 shows the genetic mapping of each GS (gs) using PCR.

Fig. 7 shows the genetic mapping of each GS (gs) using PCR.

Fig. 8 shows the genetic mapping of each GS (gs) using PCR.

Fig. 9 shows the genetic mapping of each GS (gs) using PCR.

Fig. 10 shows the genetic mapping of each GS (gs) using PCR.

Fig. 11 shows the chromosomal mapping of GS001418 (gs001418) using PCR.

Fig. 12 shows the chromosomal mapping of GS001457 (gs001457) using PCR.

Fig. 13 shows Southern blotting of human total chromosomes using the GS as a probe.

Fig. 14 shows Southern blotting of human total chromosomes using the GS as a probe.

Fig. 15 summarizes the characteristics of hybrid cells used for Southern hybridization.

Fig. 16 shows Southern blotting of chromosomal DNA from the hybrid cells using GS000152 (clone s14g02) as a probe.

Fig. 17 shows Southern blotting of chromosomal DNA from the hybrid cells using GS000041 (clone s650) as a probe.

Fig. 18 shows Southern blotting of chromosomal DNA from the hybrid cells using GS000181 (clone hm01e01) as a probe.

Fig. 19 shows Southern blotting of chromosomal DNA from the hybrid cells using GS000055 (clone c13a18) as a probe.

Fig. 20 shows Southern blotting of chromosomal DNA from the hybrid cells using GS000180 (clone s479) as a probe.

Fig. 21 shows Southern blotting of chromosomal DNA from the hybrid cells using GS000094 (clone s173) as a probe.

Fig. 22 shows Southern blotting of chromosomal DNA from the hybrid cells using junk (clone hm01g02) as a probe.

Fig. 23 shows the chromosomal mapping of each GS by Southern blotting. E stands for EcoRI, Ba stands for BamHI, Bg stands for BglII and E/B stands for double cleavage with EcoRI and BamHI.

Fig. 24 shows the chromosomal mapping of each GS by Southern-blotting. E stands for EcoRI, Ba stands for BamHI, Bg stands for BglII and E/B stands for double digestion with EcoRI and BamHI.

Fig. 25 shows the chromosomal mapping of each GS by Southern blotting. E stands for EcoRI, Ba stands for BamHI, Bg stands for BglII and E/B stands for double digestion with EcoRI and BamHI.

Fig. 26 shows the chromosomal mapping of each GS by Southern blotting. E stands for EcoRI, Ba stands for BamHI, Bg stands for BglII and E/B stands for double digestion with EcoRI and BamHI.

Preferred embodiments of the invention

In the following section, there will be explained preferred embodiments of the present invention. However, the present invention will not be restricted to these preferred embodiments.

[Example 1]

Preparation of mRNA

Cytoplasmic RNA was extracted from a liver cancer cell line HepG2 (Aden., et al., Nature 282, 615-617, 1979) using standard procedures [Sambrook, J., et al., Molecular Cloning, 2nd ed. (New York: Cold Spring Harbor Laboratory), vol. 1, pp. 7.3-7.36, 1989]. Briefly, HepG2 cells grown in Dulbecco's modified Eagle medium supplemented with 10% FCS were lysed in RNA extraction buffer [0.14 M NaCl, 1.5 mM $MgCl_2$, 10 mM Tris-HCl (pH 8.6), 0.5% NP-40, 1 mM DTT, 1000 units/ml RNase inhibitor (Pharmacia)] by using a Vortex mixer for 30 sec and then left standing on ice for 5 min. Nuclei and other cell debris were precipitated by centrifuging at 12,000 g for 90 sec, and the supernatant was deproteinized with Proteinase K followed by phenol extraction. RNA was precipitated by isopropanol and rinsed with 70% ethanol. Finally, the poly A^+ fraction was collected by oligo dT column fractionation (Aviv., et al., Proc. Natl. Acad. Sci. USA 69, 1408-1412, 1972).

[Example 2]

Preparation of vector primer DNA and construction of cDNA libraries

To prepare a vector primer, pUC19 DNA amplified in JM109 cells (Yanisch-Perron, C., et al., Gene 33, 103-119, 1985) was digested with PstI to completion and a poly T-tail was added with terminal transferase (Pharmacia) to a mean length of 26. This process was monitored by the incorporation of ^{3}H-deoxythymidine triphosphate [Okayama, H., et al., Methods in Enzymology (San Diego: Academic Press), vol. 154, pp. 3-28, 1987]. The product was digested by HincII, and the resulting short fragments were eliminated by chromatography with Sepharose S-300. Then the T-tailed plasmid was purified by an oligo dA column and stored in 50% ethanol at a concentration of 1 µg/µl.

Fig. 1 shows the outline of the construction of the cDNA library. Two micrograms each of the cytoplasmic Poly A$^+$ RNA and the vector primer DNA were co-precipitated in 70% ethanol containing 0.3 M Na-acetate and the pellet was dissolved in 12 µl of distilled water. For the first strand synthesis, after heat denaturation at 76 °C for 10 min, 4 µl of 5 x reaction buffer [250 mM Tris-HCl (pH 8.3), 375 mM KCl, 15 mM $MgCl_2$], 2 µl of 0.1 M DTT and 1 µl of 10 mM each of dATP, dCTP, dGTP and dTTP were added to the sample at 37 °C. The reaction was initiated by the addition of 200 units of reverse transcriptase MMLV-H-RT (BRL), and after incubation at 37 °C for 30 min, stopped by transferring the reaction tube onto ice. For the second strand synthesis, to the aforementioned reaction mixture the following was added: 92 µl of distilled water, 32 µl of 5 x E. coli reaction buffer [100 mM Tris-HCl (pH 7.5), 20 mM $MgCl_2$, 50 mM $(NH_4)_2SO_4$, 500 mM KCl, 250 µg/ml of BSA, 750 µM βNAD], 3 µl of 10 mM each of dATP, dCTP, dGTP and dTTP, 15 units of E. coli ligase (Pharmacia), 40 units of E. coli polymerase (Pharmacia), and 1.5 units of E. Coli RNase H (Pharmacia). The reaction mixture was then incubated at 16 °C for 2 h and heated to 65 °C for 15 min. Then 20 units each of BamHI and MboI were added, and the reaction mixture was incubated at 37 °C for 1 h and heated again at 65 °C for 30 min. Finally, the sample was diluted up to 1 ml with 1 x E. coli reaction buffer, and 100 units of E. coli ligase were added. The resulting mixture was incubated at 16 °C overnight. An aliquot of this mixture was used to transform competent E. coli DH5 cells (Toyobo). Transformants were selected by ampicillin resistance. The product was named "3' MboI cDNA library".

[Example 3]

Amplification of cDNA insert by PCR

The plasmid-carrier E. coli colonies were picked into 96-well plates containing 125 µl of LB medium (Davis, R. W., et al., Advanced Bacterial Genetics. New York: Cold Spring Harbor Laboratory, 1980) in each well and incubated in a moist chamber at 37 °C for 24 h. A replica culture was made for every plate using a 96-pinned replica device (Sigma) and the master plates were stored at -80 °C for future use. After overnight incubation at 37 °C, 50 µl of the culture from each well of these replicas were transferred to polycarbonate 96-well plates (Techne). Bacteria were collected by centrifugation in an Omnispin H4211 rotor (Sorvall) at 1500 rpm for 5 min, resuspended in 50 µl of water, covered with a layer of mineral oil and lysed at 95 °C for 30 min in a metal bath. Debris were removed by centrifugation at 3600 rpm for 30 min in the same rotor.

Five microliters of the supernatant were added to 20 µl of distilled water and kept at 95 °C for 10 min under a layer of mineral oil. Then the denatured lysate was subjected to PCR by adding 25 µl of 2 x reaction mixture [40 mM Tris-HCl (pH 8.9 at 23 °C), 3 mM $MgCl_2$, 50 mM KCl, 200 µg gelatin/ml] containing 5 pmol each of primers, 5 nmol each of dATP, dCTP, dGTP, dTTP and 1.25 units of Taq DNA polymerase (Cetus) at 70 °C. Temperature cycling reactions were carried out immediately after addition of the reaction mixtures using a thermal cycler either for microfuge tubes (PJ1000, Perkin Elmer Cetus) or for a 96-well plate (PHC-3, Techne); 35 repeated cycles of 30 sec at 96 °C, 1 min at 55 °C, and 2 min at 72 °C without a final extension step were performed.

For this method, the correct choice of primers for the PCR reaction is crucial. Therefore, preliminary tests were performed using the following primers with a predicted Tm of above 60 °C.

The primers tested were a pair of primers, FW(-47) and RV(-48), which are identical to the commercially available 24 mer primers, a second pair of primers, [FW(-40) and RV(-29)], which are a longer version (21 mer) of the well-tested sequencing primers, and the primers RV(-71) and RV(-14), which have a triplet sequence at the 3' terminus identical with that in FW(-40) but is in the opposite orientation (Fig. 2A).

In most of the cases where various combinations of primers were tested, short PCR artifacts appeared, besides the expected major products (Fig. 2B, arrows indicate the PCR artifacts.). These artifacts could be reduced by raising the annealing temperature, lowering the primer concentration or lowering the substrate

concentration but in all cases the yield of the products was not high enough to serve as a template for the sequencing reaction without concentration thereof.

However, since one pair of primers [SW(-40) and RV(-14)] did not yield artifacts (Fig. 3), this pair was selected for further tests, and was found to give reproducible results. Similar results were obtained with randomly selected cDNA clones. Therefore, only this pair of primers SW(-40) and RV(-14) was used as the primers of the present embodiment.

[Example 4]

DNA sequencing

The PCR products were drop-dialyzed against TE [10 mM Tris-HCl (pH 8.0), 1 mM EDTA] on millipore filter (VS 0.025 μm) for 90 min while stirring. Forty-eight samples are easily applied on a single filter of 150 mm diameter. Without further purification the samples were subjected to the Cycle Sequencing protocol (Applied Biosystems, 1991) using dye labeled primers with minor modifications. For dideoxycytidine sequencing reaction, 2 μl of the dialyzed PCR reaction product (about 0.2 pmol of template DNA) were added to 3 μl of a reaction mixture containing 0.4 pmol of FAM M13 (-21) Primer (Applied Biosystems) in 160 mM Tris-HCl (pH 8.9), 40 mM $(NH_4)_2SO_4$, 10 mM $MgCl_2$, 50 μM dATP, 12.5μM dCTP, 75 μM 7-deaza-dGTP (Boehringer Mannheim Biochemicals), and 50 μM dTTP, 25μMddCTP, 0.8 unit of Taq Polymerase (Perkin Elmer Cetus), and subjected to 15 plus 15 cycles of the reaction (95 • C 30 sec, 60 • C 1 sec, 70 • C 1 min and 95 • C 30 sec, 70 • C 1 min) according to the manufacturer's recommendation in a 96-well plate using a thermal cycler (PHC-3, Techne). The three other sequencing reactions for dideoxyguanosine, dideoxyadenosine, and dideoxythymidine were performed in parallel (with TMRA, JOE, and ROX primers respectively, supplied by Applied Biosystems) in an identical fashion, except that twice the volume of all the ingredients was added to the dideoxyguanosine and dideoxythymidine reactions. Each sample, from a set of four was cooled to 4 • C, pooled, precipitated with ethanol, resuspended in 6 μl of a solution of formamide/50 mM EDTA (5/1 by v/v), loaded onto sequencing gel and analyzed by a DNA autosequencer (Model 373A Ver 1.0.1, Applied Biosystems).

[Example 5]

The frequency of appearance of each GS of the cDNA library reflects mRNA population.

To confirm that our 3'-directed regional cDNA library was a non-biased representation of the mRNA population in HepG2 cells, the inserts of four cDNA clones (EF-1α, α-1-antitrypsin, hnRNP core protein A1 and inter-α-trypsin inhibitor) from the clones redundantly obtained by random selection of cDNA were radiolabeled and used as probes in a Northern analysis of poly A^+ mRNA from the HepG2 cells. (The results are shown in Fig. 4A-D, and summarized in Fig. 4F.) The relative band intensity of the four mRNA species demonstrated that their relative ratios were 52, 24, 1 and 1.2, respectively (lane iii in Fig.4F). Then the same set of probes was used for measuring the number of colonies hybridizing with each probe in the same cDNA library of 8,800 clones (Fig. 4E).

The clonal frequencies were 307, 128, 7 and 9, or in ratio, 44, 17, 1 and 1.3, respectively (lane iv in Fig. 4F). These two estimates agreed, showing that the cDNA library used is a non-biased representation of the mRNA population. The ratio was practically unchanged when different preparations of mRNA from the same cell were tested.

Fig. 4 shows the proportionality of the composition of the 3'-directed cDNA library and of the mRNA. Fig.4A, 2 μg of poly A^+ RNA from HepG2 cells was electrophoresed in lanes 1-4 of a formamide agarose gel containing ethidium bromide (5 μg/ml) and then exposed to UV. Lane 5 is the RNA ladder (BRL) used as size markers (kb). In Fig. 4B, the filter was northern blotted using the following ^{32}P-labeled 3'-specific cDNA probes: Elongation factor-1α (lane 1), α1-antitrypsin (lane 2), HnRNP core protein A1 (lane 3), inter-α-trypsin inhibitor (lane 4). In Fig. 4C, one pmol each of the non-labeled cDNA fragments [EF-1α (lane 1), α1-antitrypsin (lane 2), HnRNP core A1 (lane 3), inter-α-trypsin inhibitor (lane 4), were electrophoresed in a 2% agarose gel, then photographed. Fig. 4D is a Southern analysis of the blotted filer from Fig. 4C, using the same set of radioactive probes. Lane 5 shows the migration pattern of the reference 1 kb ladder (BRL). Hard copies of these screen images were taken at 8 h for b, and 1 h for d. The radioactivity in each band was counted directly in a scinti-scanner (β-603; Betagen) and registered in (i) and (ii) in Fig. 4F. The observed band intensities were corrected based on the band intensities in Fig. 4D (ii in Fig. 4F), and normalized relative to the value of probe 3 (HnRNP core A1, lane iii in Fig. 4F) as 1 (iii in Fig. 4F). These values represent the relative content of each mRNA species in the original mRNA preparation. Fig. 4E

shows the results of colony hybridization of the membranes carrying 8,800 colonies of the 3'-directed cDNA library using the same set of the four radioactive probes. Positive colonies were counted and registered (iv in Fig. 4F), then normalized with the value of HnRNP core protein A1 as 1. The numbers in B, D and E in Fig. 4 represent the probe No. in Fig. 4F. Fig. 4F shows a remarkable agreement between the values of lanes (iii) and (v).

[Example 6]

Population study of the cDNA library

To analyze further the composition of the cDNA library, 7 and 10 clones were selected from the redundant (group I) and solitary (group II) sequence groups, respectively, and these inserts were used as radiolabeled probes for colony hybridization (Fig. 6). The frequencies of the colonies that hybridized with group I probes were roughly identical to those that were randomly picked and sequenced. These frequencies were about 3.5%-0.1%. Nearly 52% of the cDNA library population consisted of the redundant sequence group containing 173 species. When 8 probes from group II were tested, 18 positive colonies were identified among 26,400 colonies screened, giving an average frequency of 0.007%. Two probes did not hybridize with any of the 26,400 colonies, resulting in the average frequency of <0.004%. Thus, the average frequency of the 10 probes in group II was several orders of magnitude less than the lowest of group I.

The results are summarized in Fig. 5, showing the appearance frequencies of various DNA species in the 3'-directed HepG2 cDNA library. In Fig. 5, seven cDNA probes (a15 through tb042) were selected from the 162 identified genes in the redundant group (group I), and ten (s155 through s632) were randomly chosen from the solitary group (group II). In columns A, B and C, each one of the insert DNAs was radiolabeled and used as a probe for colony hybridization tests of 982 (A), 8,800 (B) or 26,400 colonies (C). NT indicates "not tested". The DDBJ entry names of the 17 clones listed in this table are HUM000A15, HUM000C321, HUM00TB038, HUMHM01B02, HUM0C13A04, HUMHM02D02, HUM00TB042, HUM000S155, HUM000S159, HUM000S639, HUM000S635, HUM000S170, HUM000S154, HUM000S167, HUM000S645, HUM000S647, and HUM000S632.

[Example 7]

Analyses of sequencing errors

All the sequence data presented in this specification were obtained by repeated cycles of enzymatic amplification of the plasmid inserts, followed by cycle sequencing with Taq polymerase. Sequences of 60 clones that showed data bank matches were examined for discrepancies from the data bank entries. It was found that the accuracy in the region 1-100 bp distant from the cloning site was 98.7%, indicating that the primers or probes designed with the sequence in this region could be obtained practically without any erroneous sequences or even if they contain any errors, they are functionally without problems.

[Example 8]

Mapping of GS by PCR

(cDNA sequence)

cDNA library was constructed from mRNA of DMSO treated HL60 cells. The methods for construction of the 3'-directed cDNA library and for sequence analysis of the library components are the same as described in Examples 1-4.

(PCR primer)

Primer design was performed by using the computer software OLIGO 4.0 (National Biosciences) which eliminates possible formation of inter- or intra-molecular secondary structures. In addition to the primer design, transfer of oligonucleotide sequences to the local database and synthesizer were semiautomated using a Macintosh computer linked with a network. DNA oligomers were synthesized on an automated DNA synthesizer (Model 394, Applied Biosystems) on a 40 nmol scale. The synthesized oligomers were used as

PCR primers without further purification.

〈Preparation of Genomic DNA〉

The human genomic DNA was extracted from the normal karyotype lymphoblastoid cell line GM0130b. Mouse and Chinese hamster genomic DNAs were purchased from Clontech. Monochromosomal hybrid cells utilized for mapping panel were commonly used ones which have been described previously. Briefly, chromosomes 3, 4, 9, 11, 12, 13, 15, 22 and Y were carried in human-Chinese hamster monochromosomal hybrid cells, and chromosomes 1, 2, 5, 6, 7, 8, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21 and X were carried in the human-mouse monochromosomal hybrid cells A9 series. The integrity of the hybrid cells were monitored by *in situ* hybridization.

〈Amplification by Polymerase Chain Reaction〉

PCR was performed according to standard protocols (Saiki, R. K., et al., Science 230, 1350-1354, 1985), using 10 pmol of each primer on a whole 20 μl scale reaction, with 35 thermal cycles of 30 sec at 94 • C, 60 sec at an annealing temperature, and 90 sec at 72 • C, using a Perkin-Elmer 9600 thermal cycler. Annealing temperature was determined according to the "optional annealing temperature" estimated by the Program OLIGO.

〈Analysis of the PCR Products〉

The PCR products were run on an 8% polyacrylamide non-denatured gel (Acrylamide:Bis-acrylamide = 19:1, 1 mm thick) at 300 V for 1 h, followed by staining in 90 mM Tris-borate, 2 mM EDTA buffer solution containing 0.25 μg/ml ethidium bromide for 15 min. The size of the amplification products were determined relative to the 10 bp DNA ladder (BRL). Detection of fluorescence was performed by using a laser fluorescent image analyzer (FM-BIO, Hitachi Software Engineering). The image data were transferred to a computer for analysis.

〈Results of Analysis of the PCR Products〉

Among various species of 3'-directed cDNA-GSs obtained from granulocytoid cells, 195 novel GSs which did not match the sequences deposited in Genbank Release 76 were selected and used for designing primers for the PCR. The PCR was performed with these primers using the total human genomic DNA as the template.

Among the 195 primer pairs, 191 (98%) yielded products whose size matched those expected within 5 nt. The results are summarized in Figs. 6 - 10 whose figure legends are as follows: GS, gene signature; CN, clone name; Chromosomal position, chromosome numbers to which GSs were mapped; Sequence of primers, DNA sequences of primers (Sense, sense strand; anti-sense, anti-sense strand); AT, annealing temperature; HO, Observed size of PCR products with total human genomic DNA (nt); HE, Expected size of PCR products with total human genomic DNA (nt); MO, Observed size of PCR products with mouse genomic DNA (nt); CO, Observed size of PCR products with Chinese hamster genomic DNA (nt); G, Number of "hits" of GS in the granulocytoid (DMSO treated HL60) cDNA library after analyzing altogether 1000 clones; T, Total number of "hits" of the GS after analyzing altogether 3000 clones from the three cDNA libraries of HL60 with and without induction by DMSO or TPA. Question marks ("?") indicate that the PCR products did not yield a clear band.

"M" indicates that the PCR products yielded a band which was indistinguishable from the band observed after the reaction using mouse DNA as the template. Similarly, "C" indicates that the PCR products yielded a band which was indistinguishable from the band after the reaction using Chinese hamster DNA as the template.

The overall rate of success of the PCR was 191/195 (98%), although GSs were randomly selected from the cDNA sequences, indicating that the quality of the cDNA library used in this work was reliable, and that the sequence analyses and primer designs were performed appropriately. Thus, the possible chances of failure of the PCR caused by presence of an intron(s) in the relevant cDNA sequences is negligible in working with the GS, as introns virtually do not lie in the poly A proximal 3'-region of vertebrate genes (Wilcox et al., Nucleic Acids Res. 19, 1837-1843, 1991). This is a big advantage compared to the use of partial fragmented cDNA sequences obtained from randomly primed cDNA libraries (Adams et al., Science 252, 1651-1656, 1991) or from 5'-directed cDNA libraries.

(Chromosomal assignments of GS)

The 191 primer pairs that yielded PCR products from total human DNA were used for chromosomal assignments of the GSs with the monochromosomal hybrid cell panel. At least 119 GSs were assigned to a single chromosome. As an example, GS001418, shown in Fig. 11, was assigned to chromosome number 3. With some clones, extra products were obtained, some of which were assigned to the same chromosome, whereas others to different chromosomes. An example, GS001457, is shown in Fig. 12. Sixty-two (33%) clones yielded the expected PCR products with two or more different chromosomes. Thirty-five cases (18%) yielded PCR products whose size were indistinguishable from background rodent genomic DNA. Among these, 21 GSs produced products indistinguishable from mouse and Chinese hamster DNA. Ten GSs yielded no expected PCR products with the monochromosomal cell panel DNA although the expected PCR products from total human genomic DNA were observed. The 10 cases probably arose from a small deletion in the hybrid cells. Five clones obtained from HepG2 cDNA library have been analyzed also by Southern blot analysis. Four out of the 5 GSs (GS000053, GS000120, GS000271 and GS000279) gave consistent results with those obtained by the PCR. One GS (GS000228), which was uncertainly assigned to chromosome Y because of the weak signal detected by the Southern blot method, was assigned to chromosome 11 by PCR.

[Example 9]

Mapping of GS by Southern blot method

(Cell lines)

Total human genomic DNA was isolated from the human normal karyotype lymphoblastoid cell line GM0130b. Monochromosomal hybrid cells used as the mapping panel are shown in Fig. 15. Hybrid A9(neo-x)-y cells as described by Koi, et al. (Jpn. J. Cancer Res. 80, 413-418, 1989) were donated by Dr. M. Oshimura, Faculty of Medicine, Tottori University, passaged 3 times and frozen for storage. The loss or rearrangements of chromosomes could have occurred during this period. The GM series was obtained from the Mutant Cell Repository, National Institute of General Medical Science (NIGMS) (Camden, NJ). To confirm that human chromosomes remained intact in the hybrid cells after storage in liquid nitrogen, metaphase spreads of the hybrid cells were monitored by chromosome staining based on *in situ* hybridization using biotinylated total human DNA as the probe (Durnam, D. M., et al., Somatic cell Mol. Geneta. 11, 571-577, 1985) intact, as well as translocated or fragmented human chromosomes were easily detected by this means. In a hybrid cell mapping panel, chromosomes 11, 12 and 15 were represented by the hybrid cell lines A9(neo-11)-1, A9(neo-12)-4 and A9(neo-15)-2, respectively, and in another panel, they were represented by the hybrid cell lines GM10927A, GM10868 and GM11418, respectively.

(Isolation of genomic DNA and Southern blotting)

High molecular weight DNA was extracted from cells using sodium dodecyl sulfate (SDS) and Proteinase K, followed by phenol-chloroform extraction and ethanol precipitation. DNAs were digested overnight with a combination of two restriction enzymes including EcoRI, BamHI and BglII. About 5 μg of each digest was electrophoresed in an 0.8% agarose gel, then transferred to Hybond N^+ membrane (Amersham) with 0.4 N NaOH. The membrane was rinsed in 2 x SSC and stored at 4 °C for subsequent use.

Clones containing a novel sequence and having more than 150 bp were selected as probes. The cDNA inserts of the clones were amplified by the PCR. The PCR products were isolated by electrophoresis through a 2% low-melting temperature agarose gel (Nusieve : SeaPlaque, 3 : 1), followed by excision. The gel was removed by melting at 65 °C and digesting with β-Agarose I (Bio Labs) at 40 °C for 1 h. The probes were labeled with [α-^{32}P]dCTP by random priming using a commercial kit (Amersham). Hybridization proceeded at 65 °C in a high salt buffer containing 6xSSC, 1x Denhardt's solution and 0.5% SDS, in the presence of 0.1 mg/ml of sonicated, denatured salmon sperm DNA. The membranes were washed in 2xSSC, 0.1% SDS at 65 °C for 30 min, then twice for 30 min in 0.1xSSC, 0.1% SDS at 65 °C, and analyzed using a Fuji BAS-2000 imaging analyzer.

〈Analyses with Genomic DNA〉

Among the HepG2 3'-directed cDNA libraries described in Examples 1 and 2, 160 novel clones were selected and used as probes for Southern blots.

Total human genomic DNA was isolated from a cell line GM0130b that has a normal karyotype, and digested with the restriction enzymes, EcoRI, BamHI and BglII alone or in combination. The GS clones used as probes were the 3'-directed cDNAs. Each of these cDNAs covers a region between the poly(A) site and the nearest MboI site (GATC) (Okubo, K., et al., Nature Genetics 2, 173-179, 1992) and thus do not have restriction sites for BamHI or BglII. In addition, because the average size of GS is 270 bp, the chances of having an EcoRI site in the cDNA moiety were not high. In fact, only 7 clones out of the 160 analyzed had an EcoRI restriction site.

Membranes blotted with digested human genomic DNA were hybridized with radio-labeled GS probes and washed at high stringency. Since the 3'-terminal region of cDNA has, in general, a unique sequence which differs from that of protein encoding regions which tend to have conserved motifs, cross hybridization with unrelated cDNA sequences will not occur under such stringency. Examples of the results of hybridization are shown in Figs. 13 and 14. Clones s503 and s632 (Figs. 13a and 13b; junk) respectively represent unique single band producers. As shown below, 67 clones belonged to this class. The positions of the GS sequence relative to the restriction sites were inferred from the band patterns. Clone s311 (Fig. 13c; GS000092) showed a single band with EcoRI -as well as (EcoRI + BamHI)-digested DNA, but two bands of different sizes in other double digests. The double digestion thus helped resolve multiple GSs. Similar results were obtained with clone c13a08 (Fig. 13d; GS000055), in which there were 2 bands with EcoRI- or (EcoRI + BamHI)-digested DNAs, and 4 when digested with (EcoRI + BglII) or (BamHI + BglII). On the other hand, 4 hybridization bands appeared with clone s479 with EcoRI alone, but the number of bands decreased with (EcoRI + BglII) and (BamHI + BglII) (Fig. 14e; GS000180). These results indicate that genomic DNAs should be digested in various ways to reveal the maximum number of hybridizing fragments. The results of the analysis showed that 41, 10, 7 and 19 clones contained 2, 3, 4 and 5 or more bands, respectively. Clones s14f01 and tw1-46 (Figs. 14f and 14g; GS000407 and junk, respectively) contained at least 10 bands in each lane. Since the EcoRI restriction site is not present in the two GS sequences, the multiplicity of bands is likely to represent the multiple copy number of these genes. Clone kmb07 moved as a smear (Fig. 14h; junk), even after intensive high stringency washes, suggesting that this probe has a repetitious sequence which has not been hitherto identified.

〈Chromosomal assignments〉

A set of monochromosomal hybrid cells carrying a single human chromosome in a background of rodent chromosome was collected (Fig. 15). Thirteen cell lines were microcell hybrids established by Koi et al. (Koi, M., et al., Jpn. J. Cancer Res. 80, 413-418, 1989) and the others were obtained from NIGMS. The results of monitoring the human chromosomes in these cell lines by *in situ* hybridization using biotinylated total human DNA are also presented in Fig. 15.

The GSs were assigned to chromosomes using hybrid cell mapping panels. Three types of membranes were prepared, each having DNAs prepared from hybrid cells, and digested with EcoRI, (EcoRI + BamHI), or (BamHI + BglII). Among these three types of membranes, the one which should have yielded the maximum number of bands was used for each GS probe, according to the results of total genomic Southern blots. Examples of hybridization results are shown in Figs. 16 - 22. The numeral on each lane represents the human chromosome numbers which is contained in the hybrid cell, and H stands for the total human chromosomes. Clone s14g02 (GS000152; Fig. 16) that showed a single hybridization band with the total human DNA digested with EcoRI (lane H), showed the corresponding band only with the hybrid cell line containing human chromosome 4. Thus, this GS lies in chromosome 4.

The clone s650 (GS000041; Fig. 17) was assigned to chromosome 12 which showed a characteristic 7.5kb band in the presence of an (EcoRI + BamHI)-digested membrane. However, with an EcoRI digested DNA, the clone could not be assigned, as the human-specific and the cross-reacting rodent DNA fragments overlapped. The single, but shorter fragment band (1.3kb) which appeared in lanes 3, 4, 9, 13 and 22 represents the homologous DNA sequence in Chinese hamster, and the 3.3kb band in other lanes represents the homologous DNA in the mouse.

Clone hm01e01 (GS000181; Fig. 18) exhibited two fragments when hybridized to total human DNA treated with EcoRI alone, and these corresponding bands appeared in lanes 1 and 2. Thus, the two members of this gene family are located on two chromosomes.

Fig. 19 shows that clone c13a08 (GS000055) exhibited 4 bands when hybridized to (BamHI + BglII)- or (EcoRI + BglII)-digested total human DNA, although only 2 bands appeared with EcoRI- or (EcoRI + BamHI)-digested human DNA. Therefore, the (BamHI + BglII)-digested DNA panel was used for this clone. Two bands (12.3kb and 7.5kb) appeared in lane 7, a 5.2kb band in lane 2, and a 3.2kb band in lane 17. Two bands (6.0kb and 3.8kb) that cross-reacted with Chinese hamster DNA appeared in lanes 3, 4, 9, 13 and 22, and a single band (3.5kb) that cross-reacted with mouse DNA appeared in other lanes.

Clone s479 (GS000180; Fig. 20) showed 4 EcoRI fragments with total human DNA. The hybridization to an EcoRI-digested DNA panel yielded in bands of 10.5kb in lanes 7 and 19, 8.5kb in lane 8, 7.8kb in lanes 11 and 12, and 3.5kb in lane 11. Thus, the human specific genes are dispersed among chromosomes 7, 8, 11, 12 and 19, among which the 10.5 and 7.8kb bands in the total DNA both consist of two overlapping fragments. As shown in lane H, the intensity of these overlapping fragments was higher than normal. The 3.5kb band in lane H, as well as in lane 11 was also intense, suggesting that it also represents overlapping fragments.

Clone s173 (GS000094) exhibited 5 bands in EcoRI-cleaved total DNA (Fig. 21). Four corresponding fragments included a 4.5kb fragment in lane 1. Another 4.5kb band was observed in lane 4, indicating that the corresponding band in lane H overlapped. In addition, an intense 3.1kb band was observed in lane 17.

Clone hm01g02 (junk; Fig. 22) exhibited many bands with total DNA, and with those from monochromosomal hybrids. This clone must represent a multiple and closely related family of genes. It also contains a sequence conserved in homologous rodent genes which also give rise to multiple bands. Since most of the human specific and rodent bands overlapped, the chromosomes could not be assigned. Other combinations of restriction enzymes did not resolve the overlap.

The results of the total genomic DNA analyses and the chromosome assignments of 160 GSs are summarized in Figs. 23 - 26. Through total genomic DNA analyses using 4 differently digested human DNAs, 67 clones were categorized into a single band group, 41 in a two band group, 10 in a three band group, 7 in a four band group and 19 in a group that yielded five or more bands. Nine clones did not show any hybridization band under fixed conditions.

Assignment of two band clones showed that the two genes lie in different chromosomes in 15 of them, whereas the gene represented by clone s317 originated from the same chromosome. The three band clones s308 (GS000412) and s401 (GS000224) showed that two of the fragments lie on the same chromosome, and clone hm05g02 (GS000209) and s17a10 (GS000294) showed bands in different chromosomes. Clones displaying four or more bands showed a relatively dispersed distribution among chromosomes. "junk" in Example 9 is the DNA segment cloned by the same method used for GS but is not numbered.

[Example 10 Cloning of gene using GS]

[10A. Cloning of a full length cDNA encoding a human ribosomal protein, homologue of yeast S28. Cloning of the full length cDNA by PCR using a primer comprising a partial sequence of a GS(1)]

Using a primer ('5-TGAAAATTTATTACTACAGTGTTTTCACCA-3' (SEQ ID NO:7839)) that is a partial sequence of a DNA which is substantially the same as the complementary strand of HUMGS00500 and a primer (5'-TAATACGACTCACTATAGGG-3' (SEQ ID NO: 7840)) complementary to the vector (pSPORT) sequence that is located external to the 5' end of the cDNA, HepG2 cDNA library was amplified by the PCR and a full length cDNA clone encoding a human ribosomal protein, a homologue of yeast ribosomal protein S28 was isolated. (Hori et al., Nucl. Acids Res. 21: 4394, 1993).

[10B. A human ribosomal protein homologous to rat L9 ribosomal protein-Cloning of the full length cDNA by PCR using a primer comprising a partial sequence of a GS(2)]

Using a primer 5'-CTTCTTTCTGTAGCCAGGTAACTCT-3' (SEQ ID NO: 7841) that is a partial sequence of a DNA which is substantially the same as the complementary strand of HUMGS00418 and a primer (SEQ ID NO: 7840) complementary to the vector (pSPORT) sequence that is located external to the 5' end of the cDNA, a full length cDNA clone encoding a human ribosomal protein homologous to rat L9 was isolated (Hori et al., Nucl. Acids Res. 21:4395, 1993).

[10C. A human protein homologous to bovine phosphatidylethanolamine-binding protein. Cloning of the full length cDNA by hybridization using a probe comprising a partial sequence of a GS]

By hybridization with the probe,

5'-GATCGTTCTTCATGGGGGTAAGAAAAGCTGGTCTGGAGTTGCTGAATG

TTGCATTAATTGTCCTGTTTGCTTGTAGTTGAATAAAAATAGAAACCTGAAT

GAAGGAAA-3' (SEQ ID NO:7838),

that comprises a partial sequence of HUMGS00421, a full length cDNA clone encoding a human protein homologous to bovine phosphatidylethanolamine-binding protein was isolated (Hori et al., Gene 140:293, 1994).

[10D. Human mpl-ligand. Cloning of a cDNA coding for the human mpl-ligand using a GS]

This embodiment employs the 5' SLIC (single ligation to single stranded cDNA) method which is an improved version of the 5'RACE (rapid amplification of cDNA ends) method, and is described in Nucleic Acids Res., 19, 5227-5232 (1991).

① Reverse transcription of cDNA and attachment of anchor

The template was prepared using the reagents of the 5'-Amplifinder™ Kit (Toyobo, Inc.) in accordance with the protocol included therewith. Specifically, 2μg of human fetal liver poly A$^+$RNA (Clontech Laboratories, Inc.) and 10 pmol of the primer PA-6, a primer corresponding to the 3' end of the gene signature (GS) sequence HUMGS02342 and consisting of the sequence 5'-TTTTCGGCGCTCCCATTTATTCCTT-3' (SEQ ID NO: 7842), were mixed together and then denatured by heating the mixture at 65°C for 5 min. The cDNA was synthesized by combining the denatured sample with AMW reverse transcriptase, RNase inhibitor, dNTPs, and a reaction buffer, and then heating the resultant mixture at 52°C for 30 min. EDTA was then added to the mixture to stop the reaction. Thereafter, the RNA was hydrolyzed by adding NaOH to the reaction mixture and heating the resultant mixture at 65°C for 30 min. The mixture was then neutralized with acetic acid. A suspension of glass beads (GENO-BIND™) and NaI were added to the neutralized solution and the cDNA was adsorbed onto the beads. The cDNA, adsorbed onto the beads, was washed with an aqueous solution of 80% EtOH, and then eluted in 50 μl of distilled water. Glycogen was added to the solution of purified cDNA, and the cDNA was precipitated with EtOH and resuspended in 6 μl of distilled water. The resultant suspension (2.5 μl) was added to a solution containing 4 pmol of AmpliFINDER Anchor (5'-CACGAATTCACTATCGATTCTGGAACCTTCAGAGG NH_2-3') (SEQ ID NO: 7843) provided with the Kit, T4 RNA ligase, and a ligation (reaction) buffer. The reaction mixture was incubated at room temperature overnight, and the AmpliFINDER Anchor primer in the reaction mixture was thereby ligated to the 3' end of the cDNA. The ligated product was then used as a template for the subsequent PCR.

② Amplification by PCR

The primary PCR was carried out using the template produced in the procedure described above (①), the Anchor primer, 5'-CTGGTTCGGCCCACCTCTGAAGGTTCCAGAATCGATAG-3' (SEQ ID NO: 7846) and the PA-5 primer consisting of the sequence 5'-CTCGCTCGCCCATCCTTATACAGGCTCAGTTTTGTCT-3' (SEQ ID NO: 7844). Specifically, 1 μl of the template was mixed with Taq DNA polymerase (Takara Shuzo Inc., Code No. R001A), dNTPs, a PCR buffer, and 10 pmol each of the PA-5 primer and Anchor primer. The resultant reaction mixture was diluted with distilled water to a final volume of 50 μl and the PCR was performed in a DNA Thermal Cycler 480 (Perkin Elmer Cetus Corp.). The reaction mixture was subjected to 40 cycles of the PCR, wherein each cycle consisted of incubating the sample in sequence at 94°C for 1 min, 63°C for 1 min, and 72°C for 3 min and, in the last PCR cycle, at 72°C for an additional 8 min. The products of the PCR were resolved by electrophoresis in a 1% agarose gel and a broad band of

approximately 800 bp in length, representing a product of the PCR, was detected. The detected band was excised from the agarose gel and the DNA contained therein was recovered using a Sephaglas Bandprep Kit™ (Pharmacia Corp.) in accordance with the protocol included therewith. Specifically, the gel was dissolved in a solution of NaI and the resultant mixture was heated at 60 °C for 10 min. Sephaglas™ BP was added to the gel mixture and the DNA was adsorbed onto the glass beads contained therein. The glass beads, containing the adsorbed DNA, were then washed three times with a Wash Buffer provided with the Kit and eluted in 30 µl of TE buffer (10 mM Tris-HCl pH 8.0, 1mM EDTA).

One µl of the eluted DNA was used as a template in a secondary PCR. In order to enhance the specificity of the secondary PCR, the reaction was performed with PA-4 primer which consisted of the sequence 5'-CTCGCTCGCCCATGTATAGGGACAGCATTTCTGAGAG-3' (SEQ ID NO: 7845) and was positioned within the template sequence internal to the PA-5 primer and the Anchor primer. Specifically, 1 µl of the template was mixed with 2.5 units of Taq DNA polymerase (Takara Shuzo Inc., Code No. R001A), dNTPs, a PCR buffer, and 10 pmol each of the PA-4 primer and Anchor primer. The resultant reaction mixture was diluted with distilled water to a final volume of 50 µl preheated at 94 °C for 6 min, and the secondary PCR was then performed under the same conditions described above (①) for the primary PCR. The products of the secondary PCR were resolved by electrophoresis in a 1% agarose gel and a broad band of approximately 800 bp in length, representing a product of the PCR, was detected . The detected band was excised from the agarose gel and the DNA contained therein was recovered and purified under the same conditions as described above (①) for the primary PCR.

③ Subcloning into plasmid vector

The purified DNA product of the secondary PCR was subcloned into the plasmid vector pUC18 (pharmacia Corp.), using a SureClone™ Ligation Kit (Pharmacia Corp.) in accordance with the protocol included therewith. Specifically, the purified DNA was added to a solution containing Klenow polymerase, polynucleotide kinase and a reaction buffer, mixed and heated at 37 °C for 30 min in order to create blunt-ended termini and to phosphorylate the 5' terminus of the DNA molecules contained in the reaction mixture. The blunt-ended and phosphorylated DNA was combined with a solution containing 50 ng of a dephosphorylated and Sma I-cleaved pUC18 vector provided with the Ligation Kit, T4 DNA ligase, DTT and a ligation reaction buffer, and the resultant mixture was warmed at 16 °C for 3 hr. One sixth volume of the reaction solution was employed to transform E. coli competent cells using standard methods. Specifically frozen E. coli competent cells (Wako Pure Chemical Industries, Ltd.) were thawed and mixed with the ligated DNA. The resultant mixture was incubated on ice for 20 min, heat-treated at 42 °C for 45 sec, and then incubated on ice for 2 min. A medium [HI-Competence Broth (Wako Pure Chemical Industries, Ltd.)] was added to the mixture containing the transformed E. coli cells. The mixture was incubated for 37 °C for 1 hr and then spread onto agar plates containing 100 µg/ml Ampicillin, 40 µg/ml X-Gal (6-bromo-4-chloro-3-indolyl-β-D-galactoside), 0.1 mM IPTG (isopropyl-β-D-thiogalactopyranoside) and cultured overnight at 37 °C. White colonies were selected from the colonies which consequently appeared on the agar plates and analyzed by the PCR to determine the presence or absence of the DNA insert. Specifically, a sample of a selected colony was picked with a sterilized toothpick and used to inoculate a 50 µl reaction solution containing 1 unit of Taq DNA polymerase, dNTPs, PCR buffer, 200 µM each of the M13 P4-22 primer consisting of the sequence 5'-CCAGGGTTTTCCCAGTCACGAC-3' (SEQ ID No: 7847) and M13 P5-22 primer consisting of the sequence 5'-TCACACAGGAAACAGCTATGAC-3' (SEQ ID No: 7848), wherein both primers are comprised of sequences complementary to the pUC18 vector. The resultant mixture was heated at 94 °C for 6 min and then subjected to 30 cycles of the PCR wherein each cycle consisted of incubating the sample in sequence, at 94 °C for 1 min, 55 °C for 1 min, and 72 °C for 2 min. The amplified insert was detected by electrophoresis of the PCR products on an agarose gel and thereby the clone pR02342-2, containing an insert, was selected.

④ Sequencing of cDNA

The plasmid DNA was prepared using the QIAPrep-Spin Kit (Funakoshi, Ltd.) in accordance with the standard alkali-SDS protocol included therewith. Specifically, E. coli cells transformed with the DNA of clone pR02342-2 were cultured overnight in Luria Broth medium containing 100 µg/ml Ampicillin. The cultured cells were then pelleted by centrifugation and resuspended in P1 solution provided in the Kit. The resultant cell suspension was mixed with the P2 alkali solution of the Kit, incubated at room temperature for 5 min, neutralized with N3 solution of the Kit, incubated on ice for an additional 5 min and then centrifuged. The supernatant obtained from the centrifuged solution was applied to a QIAPrep-Spin column. The Spin column

was then washed in sequence with PB and then PE solution of the Kit and the DNA was eluted from the column with TE buffer. Sequencing of the eluted DNA was then carried out using the sequencing kit PRISM™ Terminator Mix (Applied Biosystem Corp). Approximately 1 μg of the purified DNA was mixed with a solution containing 3.3 pmol of either the M13 P4-22 primer or M13 P5-22 primer and 9.5 μl of PRISM™ Terminator Mix. The M13 P4-22 and M13 P5-22 primer were used to sequence both strands of the DNA insert of clone pR02342-2. The resultant mixture was diluted to a final volume of 20 μl with distilled water and subjected to 25 cycles of the PCR wherein each cycle consisted of incubating the sample in sequence at 96 °C for 30 sec, 50 °C for 15 sec, and 60 °C for 4 min. The excess primers and fluorescent dye present in the reaction mixture were removed by gel filtration using a MicroSpin™ S-200 HR column (Pharmacia Corp.) and the DNA products of the sequencing reaction were precipitated with EtOH. The precipitated DNA was resuspended, sequenced using an automated sequencer, "Model 373A" (Applied Biosystem Corp.), and thereafter analyzed to determine the nucleotide sequence.

The analysis of the nucleotide sequence revealed that the insert of clone pR02342-2, including the PA-4 primer, was 608 bp in length. The sequence of this insert was subjected to a search for homologous sequences entered in the Gen Bank data base, and a 100% match was found to a sequence in the cDNA which encodes the human mpl-ligand (Accession No. L 33410, Nature 369, 533-538, 1994). Further comparison of the insert of clone pR02342-2 with the cDNA sequence of the human mpl-ligand revealed that the cloned insert contained 81 bp of the 3' coding region of open reading frame. In addition, the insert of clone pR02342-2 contained an additional sequence extending beyond the 3' end of the human mpl-ligand cDNA sequence registered under Gen Bank Accession No. L 33410. These findings suggest that, using the GS HUMGS02342, the inventors of the present invention succeeded in cloning a cDNA clone pR02342-2, which could possibly have a different and more desirable property for expression than the human mpl-ligand cDNA represented by the sequence registered under Gen Bank Accession No. L 33410.

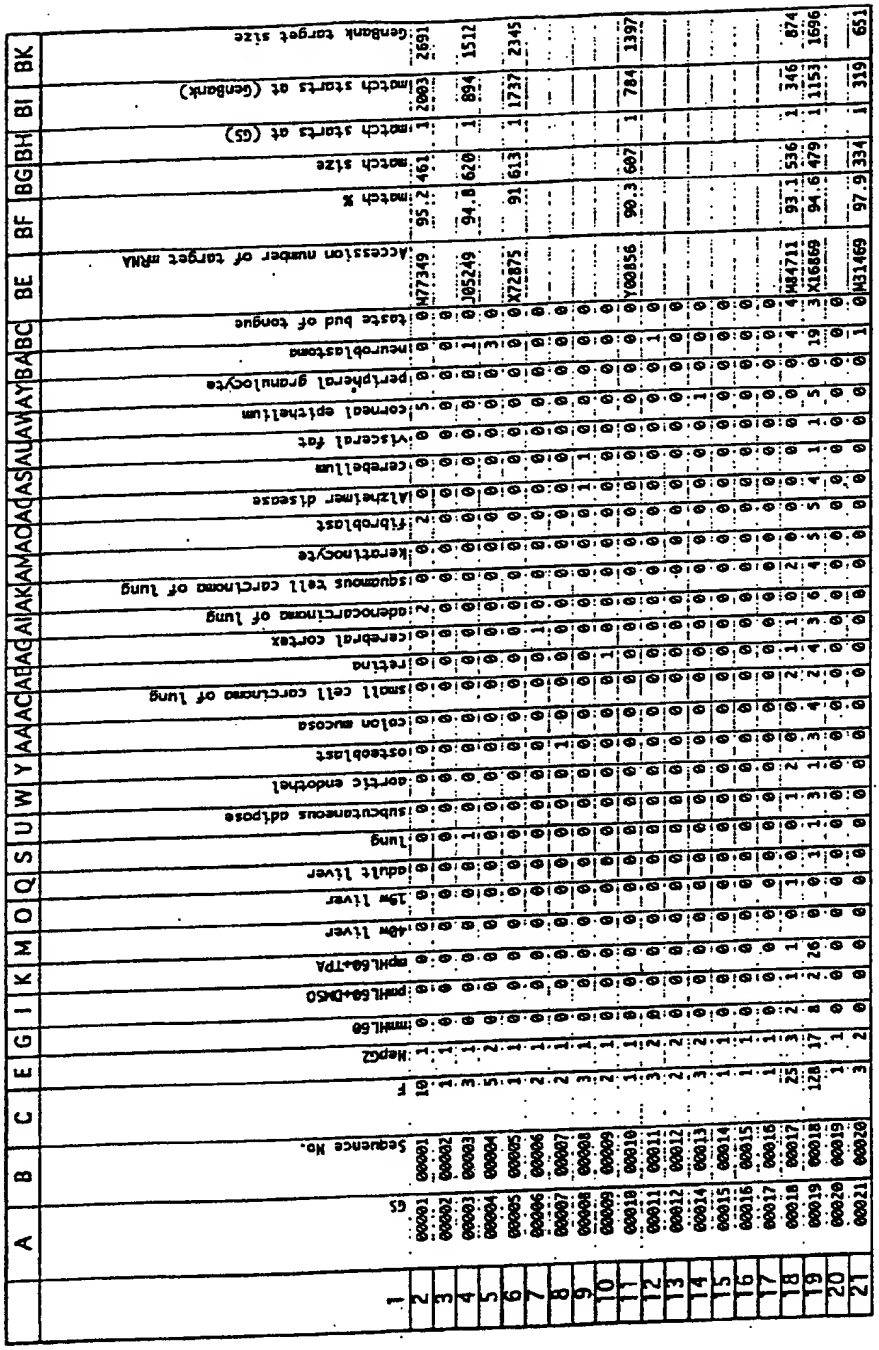
⑤ Cloning of the full-length cDNA encoding the human mpl-ligand

In order to find an optimal PCR primer, an appropriate computer program is used to search the sequence downstream of the coding region of the human mpl-ligand (clone pR02342-2) and thereby a primer PA-7 is designed and synthesized. A PCR similar to that described above in ② is performed using the template produced by the procedure described above in ①, the Anchor primer, and the PA-7 primer. Specifically, 1 μl of the template is mixed with 2.5 units of Taq DNA polymerase (Takara Shuzo Inc., Code No. R001A), dNTPs, a PCR buffer, and 10 pmol each of the PA-7 primer and Anchor primer. The resultant reaction mixture is diluted with distilled water to a final volume of 50 μl and the PCR is performed in a DNA Thermal Cycler 480 (Perkin Elmer Cetus Corp.) under conditions similar to that described above in ②. The products of the PCR are then resolved by electrophoresis on a 1% agarose gel and a band greater than 1300 bp in length, representing a product of the PCR, is recovered and cloned into a suitable vector in a manner similar to that described in ③. The cloned DNA is sequenced in a manner similar to that described in ④. The sequence is then compared to that of the human mpl-ligand cDNA registered under Gen Bank Accession No. L 33410 to confirm the presence of the full-length open reading frame.

Alternatively, using the Takara La PCR Kit (Takara Shuzo Inc., Code No. RR011) in accordance with the protocol included therewith, performing the 5'RACE procedure using primers similar to those described above in ②, a cDNA of approximately 2 Kb in length, corresponding to the human mpl-ligand, was isolated.

The tables of appearance frequencies for all GSs related to the present invention are followed by "Sequence Listing" for these GSs, wherein HUMGS numbers after the heading 'clone' represent GS numbers. In the sequence table, N in the base sequence stands for "A or C or G or T or U". However, since nucleic acids in the Sequence Listing are DNAs, "T or U" stands for T in this case.

By the present invention, it has become possible to provide DNA molecules which carry "the information for expression" in various cells and can be used for detecting and diagnosing the cellular abnormalities, recognizing and identifying cells and further efficiently cloning genes which are expressed in a tissue-specific manner, and furthermore cloned DNA molecules which can be used for the production of proteins useful as pharmaceutical products.



	A	B	C	E	G	I	K	M	O	Q	S	U	W	Y	AA	AC	AE	AG	AI	AK	AM	AO	AQ	AS	AU	AW	AY	BA	BC	BE	BF	BG	BH	BI	BK
	GS	Sequence No.		F	NeprG2	mmpL60	pmpL60+DR50	mpHL60+TPA	40w liver	19w liver	adult liver	lung	subcutaneous adipose	aortic endothel	osteoblast	colon mucosa	small cell carcinoma of lung	retina	cerebral cortex	adenocarcinoma of lung	squamous cell carcinoma of lung	keratinocyte	fibroblast	Alzheimer disease	cerebellum	visceral fat	corneal epithelium	peripheral granulocyte	neuroblastoma	taste bud of tongue	Accession number of target mRNA	match %	match size	match starts at (GS)	match starts at (GenBank)	GenBank target size
1																																		
2	00001	00001		10	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	M77349	95.2	461	1	2003	2691
3	00002	00002		1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	J05249	94.8	620	1	894	1512
4	00003	00003		3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	X72875	91	613	1	1737	2345
5	00004	00004		5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
6	00005	00005		1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
7	00006	00006		2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
8	00007	00007		2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
9	00008	00008		3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
10	00009	00009		2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	Y00856	90.3	607	1	784	1397
11	00010	00010		1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1						
12	00011	00011		3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
13	00012	00012		2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
14	00013	00013		3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
15	00014	00014		1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
16	00015	00015		1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
17	00016	00016		1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
18	00017	00017		25	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	M84711	93.1	536	1	346	874
19	00018	00018		126	17	8	2	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	X16869	94.6	479	1	1153	1696
20	00019	00019		1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19						
21	00020	00020		3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	M31469	97.9	334	1	319	651

Table 1

	A	B	C	E	G	I	K	M	O	Q	S	U	W	Y	AA	AC	AE	AG	AI	AK	AM	AO	AQ	AS	AU	AW	AY	BA	BC	BE	BF	BG	BH	BI	BK
5242	06256	05241	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5243	06257	05242	5			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	2	0	0	0					
5244	06258	05243	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5245	06259	05244	3			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	X52056	98.5	68	1	1297	1364
5246	06260	05245	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0					
5247	06261	05246	2			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0					
5248	06262	05247	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5249	06264	05248	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5250	06267	05249	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5251	06268	05250	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5252	06269	05251	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5253	06270	05252	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5254	06271	05253	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5255	06272	05254	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5256	06273	05255	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	M80359	93.9	181	1	2062	2914
5257	06274	05256	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5258	06275	05257	5			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	2	0	0	0	M90104	94.2	173	1	1700	1879
5259	06277	05258	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5260	06278	05259	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5261	06279	05260	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5262	06280	05261	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	M64784	95.6	338	1	638	1138
5263	06281	05262	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5264	06282	05263	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5265	06283	05264	11			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	X52426	100	102	1	1574	1677
5266	06284	05265	2			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0					
5267	06286	05166	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0					
5268	06288	05167	2			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5269	06289	05168	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0					
5270	06290	05169	2			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5271	06291	05170	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5272	06292	05271	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5273	06293	05272	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5274	06294	05273	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5275	06295	05274	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5276	06296	05275	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0					
5277	06297	05276	1			0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	M86409	92.4	197	1	1196	2834

Table 147

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT:
(A) NAME: CHUGAI PHARMACEUTICAL CO., LTD.
(B) STREET: 41-8, Takada 3-chrome, Toshima-ku
(C) CITY: Tokyo
(E) COUNTRY: JAPAN
(F) ZIP: 171

(ii) TITLE OF INVENTION: GENE SIGNATURE

(iii) NUMBER OF SEQUENCES: 7848

(iv) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: Diskette, 3.5 in., DS, 1.44 MB
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: PC-DOS/ MS-DOS
(D) SOFTWARE: MS-DOS

(v) CURRENT APPLICATION DATA
(A) APPLICATION NUMBER: EP 95900295.7

(vi) PRIOR APPLICATION DATA
(A) APPLICATION NUMBER: PCT/JP94/01916
(B) FILING DATE: 11. November 1994

GATCATCACT AGCAGATGTC AGTTGCACAT TGAGTCCTTT ATGAAATTCA TAAATAAAGA 60
ATTGTTCTTT CTTTGTGGTT TTAATAAGAG TTCAAGAATT GTTCAGAGTC TTGTAAATGT 120
TATTTTAATA ATCCCTTTAA ATNTNATCTG TTGCTGTTAC CTCTTGAAAT ATGATTTATT 180
TAGATTGCTA ATCCCACTCA TTCAGGAAAT GCCAGGNAGG TATTCCTTGG GGAAATGGTG 240
CCTCTTACAG TGTAAATNTT NCCTCCTGNA CCTTTGCTAA TATCATGGCA GANTTNNCCT 300
NATCCCTTTG TGAGGCAGTT TN 322

SEQ ID NO:5251
SEQUENCE LENGTH:215
SEQUENCE TYPE:nucleic acid
TOPOLOGY:linear
CLONE:HUMGS06269
SEQUENCE DESCRIPTION:
GATCAAAAAT AAGATTACAG TTAAAATATT NCTATATTCA GATGGTTTAG AGACCAGGCT 60
GTAGAATCAG ACAGCCCTGA AATTGTATCA CACANGGCTG TGTTACCCTG TACAAATAAC 120
TTAGCCTTAC TAAGCCTGTA TTTCCTCATC TGCAAANTAG GGNTGNTAAT ATACCTGTNG 180
NTAAANATGT TTTCATTAAA AANCGTTGGC GCAAA 215

SEQ ID NO:5252
SEQUENCE LENGTH:229
SEQUENCE TYPE:nucleic acid
TOPOLOGY:linear
CLONE:HUMGS06270
SEQUENCE DESCRIPTION:
GATCTAAAGT GCAGCAGAGT GGCTGNTGCT GCAAGTNATG TCTAAGGCTA GGAACTATCA 60
GGTGTCTATA ATTGTAGCAC ATGGAGAAAG CAANTGTAAA ACTGGATAAG AAAATTATTT 120
TGGCAGTTCA GCCCNTTCCC TTTTTCCCAC TAANTTTTTN CTTAAATTAC CCATGTAACC 180
ATTTNANCTC TCCAGTGCAC TTTGCCATTA ANGTCTCTGC ACATTGAAA 229

SEQ ID NO:5253
SEQUENCE LENGTH:219
SEQUENCE TYPE:nucleic acid
TOPOLOGY:linear
CLONE:HUMGS06271
SEQUENCE DESCRIPTION:
GATCGTGAAA GAGGCTTACC CAGACCACAC ACANGTTTGA GAAAAACAAT CCCCATTATN 60
ACCCATCTAG CAAAGAGGAC AACCCTAAGT GGTCCATGNG TGGATGTACA GTTTGTNCGG 120
ATGATGAAGC GTTTCATTCC CCTGGCTGAG CTCAAATCCT GTCATCAAGG CNCACAANGC 180
TACTGNTGGC CCCTNAAAAA ATATTGTTNT NTGTCACTN 219

SEQ ID NO:5254
SEQUENCE LENGTH:144
SEQUENCE TYPE:nucleic acid
TOPOLOGY:linear
CLONE:HUMGS06272

2. A DNA probe consisting of a purified single-stranded DNA , a purified single-stranded DNA complementary thereto, or a purified double-stranded DNA consisting of said single strands, containing all or a portion of a single-stranded DNA or a single-stranded DNA complementary thereto comprising any of the base sequences listed under SEQ ID NO 1-7837 and hybridizing specifically to a particular site of human genomic DNA, human cDNA or human mRNA.

3. A DNA primer consisting of a purified single-stranded DNA, a purified single-stranded DNA complementary thereto, or a purified double-stranded DNA consisting of said single strands, containing all or a portion of a single-stranded DNA or a single-stranded DNA complementary thereto comprising any of the base sequences listed under SEQ ID NO 1-7837 and hybridizing specifically to a particular site of human genomic DNA, human cDNA or human mRNA.

4. A purified single-stranded DNA, a purified single-stranded DNA complementary thereto, or a purified double-stranded DNA consisting of said single strands, containing all or a portion of a single-stranded DNA or a single-stranded DNA complementary thereto, wherein said single-stranded DNA is complementary to a human mRNA containing any of the base sequences listed under SEQ ID NO 1-7837 (wherein T is read as U) or any portion thereof at its 3' region, and hybridizing specifically to a particular site of human genomic DNA, human cDNA or human mRNA.

5. A DNA probe consisting of a purified single-stranded DNA, a purified single-stranded DNA complementary thereto, or a purified double-stranded DNA consisting of said single strands, containing all or a portion of a single-stranded DNA or a single-stranded DNA complementary thereto, wherein said single-stranded DNA is complementary to a human mRNA containing any of the base sequences listed under SEQ ID NO 1-7837 (wherein T is read as U) or any portion thereof at its 3' region, and hybridizing specifically to a particular site of human genomic DNA, human cDNA or human mRNA.

6. A DNA primer consisting of a purified single-stranded DNA, a purified single-stranded DNA complementary thereto, or a purified double-stranded DNA consisting of said single strands, containing all or a portion of a single-stranded DNA or a single-stranded DNA complementary thereto, wherein said single-stranded DNA is complementary to a human mRNA containing any of the base sequences listed under SEQ ID NO 1-7837 (wherein T is read as U) or any portion thereof at its 3' region, and hybridizing specifically to a particular site of human genomic DNA, human cDNA or human mRNA.

SEQ ID NO:7844
SEQUENCE LENGTH:37
SEQUENCE TYPE:nucleic acid
STRANDEDNESS:single
TOPOLOGY:linear
SEQUENCE DESCRIPTION:
CTCGCTCGCC CATCCTTATA CAGGCTCAGT TTTGTCT 37

SEQ ID NO:7845
SEQUENCE LENGTH:37
SEQUENCE TYPE:nucleic acid
STRANDEDNESS:single
TOPOLOGY:linear
SEQUENCE DESCRIPTION:
CTCGCTCGCC CATGTATAGG GACAGCATTT CTGAGAG 37

SEQ ID NO:7846
SEQUENCE LENGTH:38
SEQUENCE TYPE:nucleic acid
STRANDEDNESS:single
TOPOLOGY:linear
SEQUENCE DESCRIPTION:
CTGGTTCGGC CCACCTCTGA AGGTTCCAGA ATCGATAG 38

SEQ ID NO:7847
SEQUENCE LENGTH:22
SEQUENCE TYPE:nucleic acid
STRANDEDNESS:single
TOPOLOGY:linear
SEQUENCE DESCRIPTION:
CCAGGGTTTT CCCAGTCACG AC 22

SEQ ID NO:7848
SEQUENCE LENGTH:22
SEQUENCE TYPE:nucleic acid
STRANDEDNESS:single
TOPOLOGY:linear
SEQUENCE DESCRIPTION:
TCACACAGGA AACAGCTATG AC 22

Claims

1. A purified single-stranded DNA, a purified single-stranded DNA complementary thereto, or a purified double-stranded DNA consisting of said single strands, containing all or a portion of a single-stranded DNA or a single-stranded DNA complementary thereto comprising any of the base sequences listed under SEQ ID NO 1-7837 and hybridizing specifically to a particular site of human genomic DNA, human cDNA or human mRNA.

Fig. 1

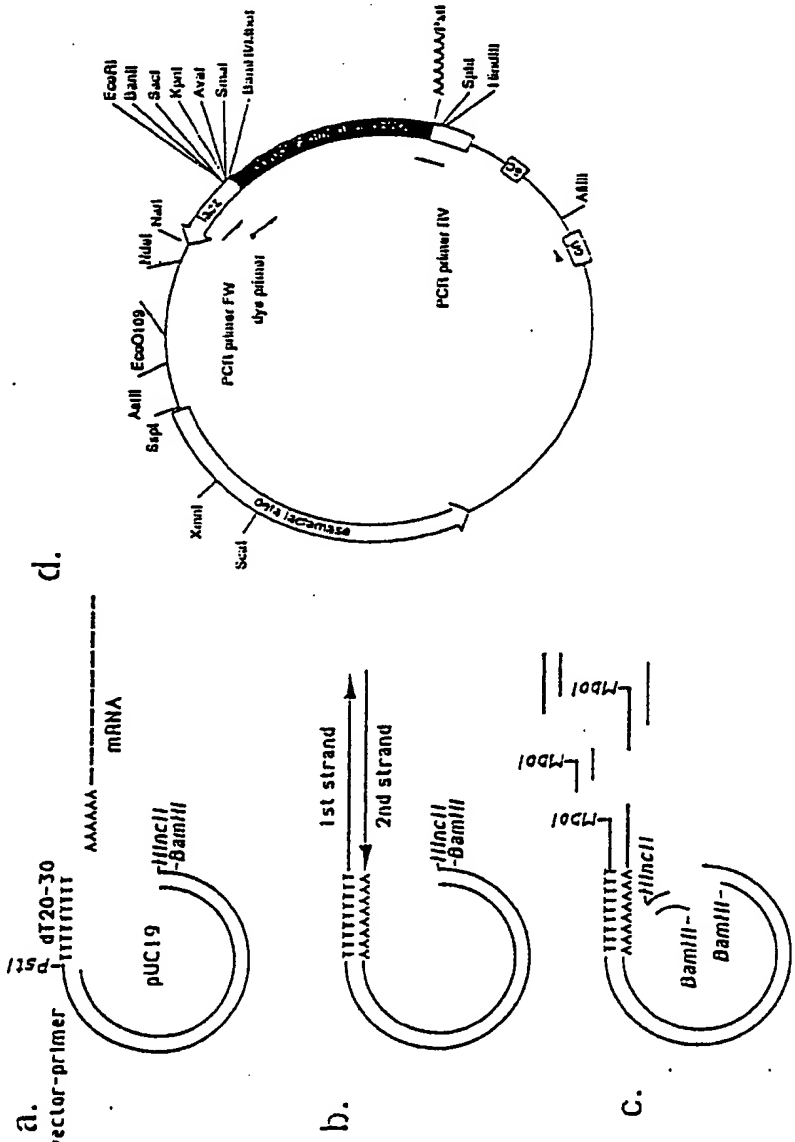

a.
vector-primer
dT20-30
pUC19
mRNA
HincII
BamHI
b.
1st strand
2nd strand
c.
BamHI-
d.
EcoRI
BanII
SacI
KpnI
AvaI
SmaI
AhdI
NaeI
EcoO109
AatII
BspI
XmnI
ScaI
PCR primer FW
dye primer
PCR primer RV
AflIII
SphI
HindIII
ori
Amp lactamase

Fig. 2

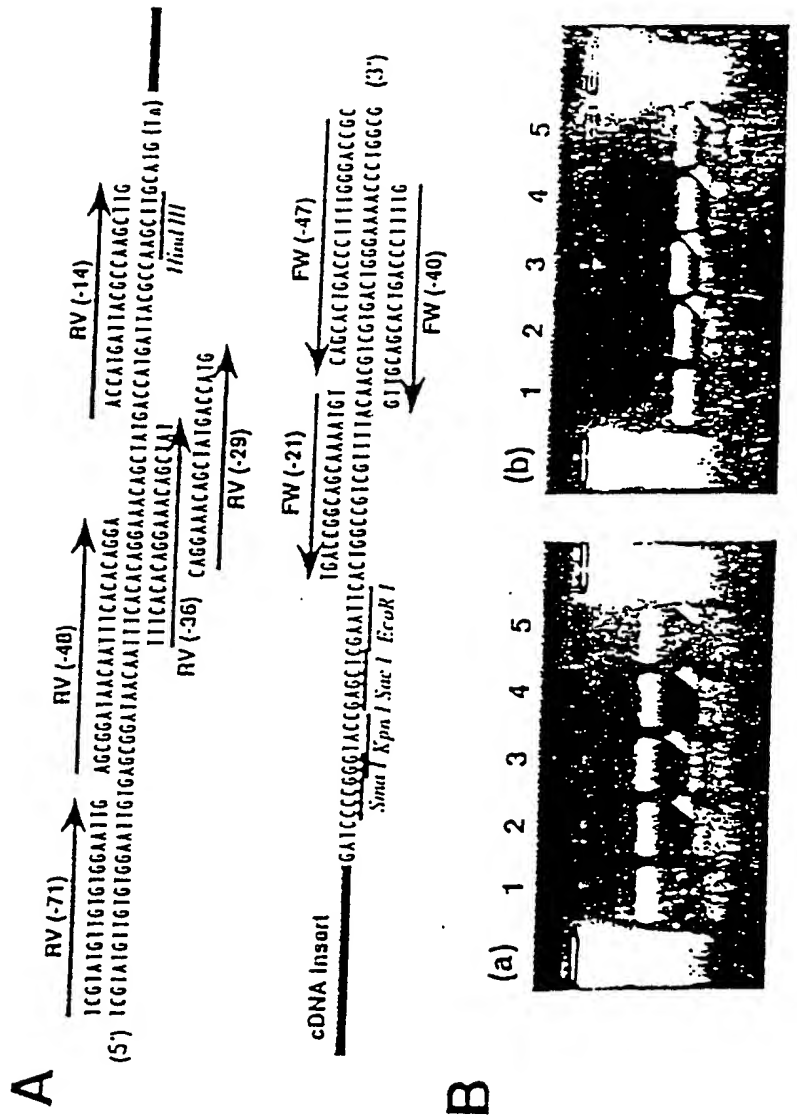

A
RV (-71)
RV (-40)
RV (-14)
RV (-36)
RV (-29)
FW (-21)
FW (-47)
FW (-40)
cDNA Insert
(5')
(3')
Sma I Kpn I Sac I EcoR I
Hind III
B
(a) 1 2 3 4 5
(b) 1 2 3 4 5

Fig. 3

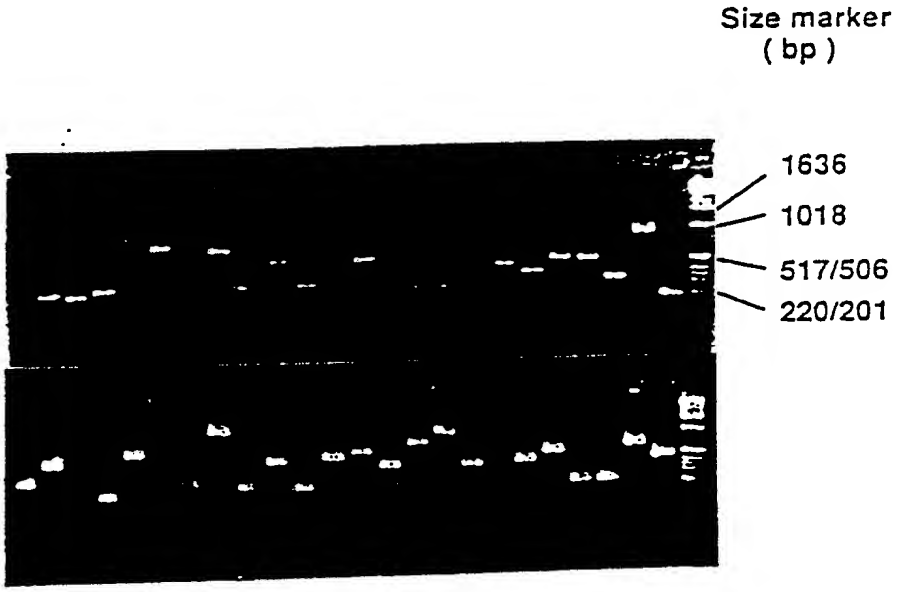


Fig. 4

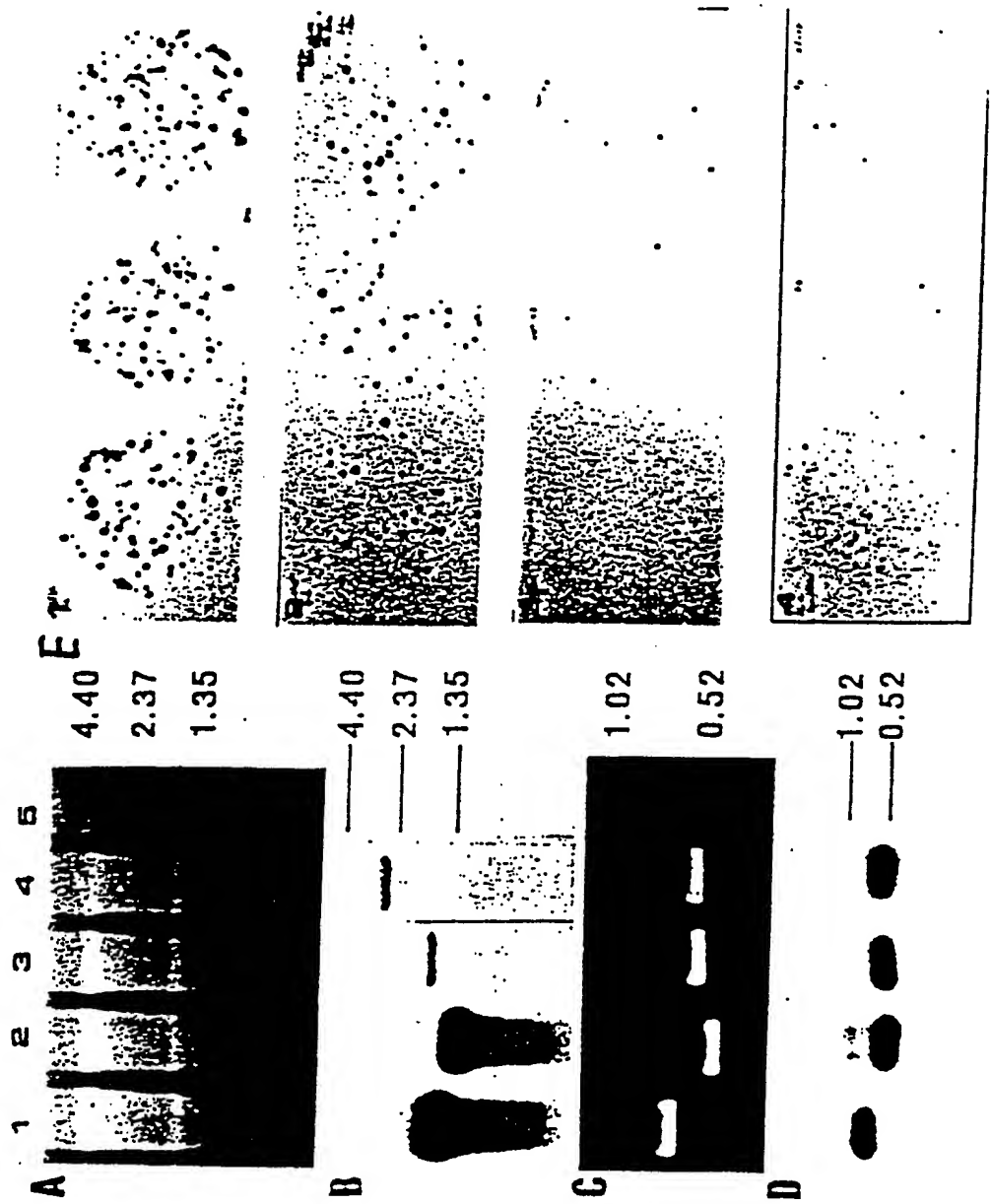


Fig. 4

F

probe No.	1	2	3	4
gene	Elongation factor 1-α	α1-antitrypsin	HnRNP core protein A1	Inter-α-trypsin inhibitor
(a) Band intensity of Northern blot(cpm)	687	423	10	15
(b)Band intensity of control blot(cpm)	133	177	100	127
(c)Normalized signal(a)/(b)x10	52	24	1	1.2
(d)Positive signals on colony blot	307	119	7	9
(e)Relative representation	44	17	1	1.3

Appearance frequencies of various cDNAs in the 3'-directed HepG2 cDNA library

Group	Clone	Gene	A In 982 (%)	B "in 8,800 (%)"	C "in 26,400 (%)"
I	a15	Elongation factor - 1Aα	22 (2.2)	307 (3.5)	NT
	c321	Translationally restricted tumor protein	12 (1.2)	89 (1.0)	NT
	lb038	α-1-antitrypsin	8 (0.8)	119 (1.4)	NT
	hm01b02	Light chain of ferritin	6 (0.6)	62 (0.7)	NT
	c13a04	NADP(H) Menadione oxidoreductase	4 (0.4)	27 (0.3)	NT
	hm02d02	Ribosomal protein S11	3 (0.3)	29 (0.3)	NT
	lb042	Human RNP core protein A1	2 (0.2)	7 (0.1)	NT
II	s155	unknown	1	2	5 (0.02)
	s159	unknown	1	2	4 (0.02)
	s639	unknown	1	1	3 (0.01)
	s635	unknown	1	0	2 (0.01)
	s170	unknown	1	0	1 (0.004)
	s154	unknown	1	0	1 (0.004)
	s167	unknown	1	0	1 (0.004)
	s645	unknown	1	0	1 (0.004)
	s647	unknown	1	0	0 (<0.004)
	s632	unknown	1	0	0 (<0.004)

Fig. 5

EP 0 679 716 A1

GS	CII	Chromosomal position	Sequences of primers		AT	HO	HE	MO	CO	G	T
			Sense	Anti-sense							
gs000788	pm2366	1	CAGAGCCCCAGTACAACTAT	AAGTTTATTGTTGGGGTCAG	48	114	115	104	110	1	2
gs001026	pm2444	1	AATGGGACAGTTACACTTGA	CCAGCTTCCTTGACTTGAGA	48	83	81	>200	>200	1	1
gs001075	pm0883	1	TGGACTGTGGATACCTATCT	ACAAGTACCCCTGAATGGCT	48	124	124	103	107	4	4
gs001087	pm1772	1	GTCACTCTCAGCCATAGCAC	ACCATCTTCAGCCCACACTT	50	104	104	180	>200	6	6
gs001094	pm0347	1	GCCCCTAACACGAGGAACTC	TAATTTCCCACTCCCGTAAC	51	114	118	>200	200	1	1
gs001116	pm1771	1	GGGTTTTCAATAGGGTGTAGACC	GCCCCAATCTGTCAAAACTG	49	95	95	78	107	1	1
gs001191	pm0609	1	TTGCTGGATTGTAACTTTTG	GGCTGAACATTCACTCTTTG	47	97	97	-	200	1	1
gs001200	pm1351	1	TTAAGAAGACCCTTATGGAGACC	AATAATCTTGGTTAGTCACTTAC	47	97	98	-	-	1	1
gs001346	pm0982	1	TCAGGTCTGCTTGGAGGATG	AACTCACAGCACAGTATTTG	53	120	122	>200	>200	1	1
gs001446	pm1518	1	AAGGTGTACAGGATATTTGCAGA	TGCAATAGCCCAATCTCATT	47	130	125	>200	>200	1	1
gs001464	pm1439	1	CCAAAGACCTCCGTTGAACA	TTTGGGAGAGCCATAGACAG	51	100	100	>200	-	1	1
gs001468	pm0427	1	TACTCAGTGGAAAGATAAAC	CAGTGGACCACATTTTCTTA	40	98	98	-	-	2	2
gs001521	pm2785	1	CCCAAATCAAATTGTTAAATG	TTTGAATCAGAGACATGAAGTT	43	102,175	100	>200	>200	1	1
gs001554	pm2291	1	CCAGAGAGTTCAAGGGATTG	GGTACAAAGTGCAAATGACT	46	57	57	78	155	1	1
gs001572	pm2006	1	CCAACATGGTCCTAGCACTG	AAACTTTATTGCAGCTTCTT	44	58	58	>200	>200	4	4
gs000120	pm1350	2	CATGATACTCTTCGGTGGTA	AAACAGTAGTTGCCAGCATT	46	84	109	-	-	1	1
gs001036	pm1730	2	AGGCTGAAATGTGGCATGCT	CCCGTTATTGCTACATGTCT	48	119	119	93	115	1	1
gs001081	pm0931	2	AAAGCAATACAAAATACCAA	TTCAATATGTTTAACCAGTA	40	90	90	-	-	1	1
gs001090	pm0925	2	TAATGTACCACGATGAATAG	TAATGTAATAATGCAGGTAA	45	88	88	-	-	1	1
gs001213	pm2010	2	CCAGATGGAAAGGGAAGTCT	CTGGAATATGGAGAATCAAACAG	47	125	125	150	>200	1	1
gs001252	pm0935	2	TCGAGTTTTGTCTCTAATAA	GGAAATAATCGCTTCAGTTG	43	103	103	-	-	1	1
gs001268	pm2093	2	AGTCCTTCTTGGCTCCTCAT	TATCGTCAGTGCCTTTATTG	52	137	137	>200	-	1	1
gs001438	pm2435	2	TTTTGTACCTACGTAAGAGTACTT	ATCCGTGCCACACATAGTGA	45	105	108	-	-	1	2
gs001442	pm1671	2	TTATTAGGGAGTCATTATTCTGTG	AGTTCCCATTCTTCCACATG	45	67	65	>200	>200	1	2
gs001453	pm1245	2	TTGCTTTCCCGTCTCTAAGT	ATGTACAATTTGCGTATGTAGG	45	75	75	170	190	1	1
gs001535	pm1246	2	ATCTACTGTTTGTTGAAGTG	ACTGATTTTGGTCCCATCTG	44	68	67	-	-	1	1
gs000875	pm0449	3	CGAACATTTCACCTCTCATA	ATGATTTATTTAGGCAGGAA	43	68	65	-	-	3	6
gs001001	pm1758	3	TCTGGCTCTTTGGTGTTGGA	GGCCCACTGAGTACAATGTC	51	115	115	-	-	1	1
gs001218	pm2424	3	AAAGAAAGCACACTGCCTAA	ATGTATAGACAAATCCAAAG	42	90	90	-	-	1	1
gs001219	pm0668	3	GTAGTCTCCTGCCCTTTAGC	AAGGATTTGATTTTCTACAT	43	77	77	-	-	1	1
gs001277	pm1729	3	GGTCCTGTTTATTTTGACAT	AAACAAGAGGATGGTTCAGA	43	75	75	155	>200	1	1
gs001306	pm1822	3	GATCCTTGGTGTGTAGTTCAGTC	CTGCAAAATACAGGGAATCAT	46	83	83	160	140	1	1
gs001418	pm2209	3	ACCCCAGTCCCAAATCCAGT	ACACTCCCCCAGCCCTTACT	55	105	105	113	>200	1	1
gs001468	pm2455	3	ATCTAGCTGGCTGTAGTATT	TTAAAGAGATGAATTTATTGGT	42	110	110	190	>200	1	1
gs000271	pm1252	4	GTCCTTTGCTATCTGTGTTA	AAGCATTTATTTGAGGTTTA	43	90	90	95	>200	1	3

Fig. 6

EP 0 679 716 A1

Fig. 7

| | | |

gs001044	pm1659	13	TTGTAAGCCTATCAGAGTCA	AGACAGACTTATGCCATCTA	44	109,200	100	>200	100	1	1
gs001290	pm1731	13	GCTTCTTCCTGTGCTGTGGT	GCAGTTAATCATGGCTATTCTCC	50	122	122	>200	150	1	1
gs001362	pm0118	13	ACTGAAATGGAACATAGTCT	TACATTACATGACATTGTGA	40	61	61	95	103	1	1
gs001366	pm0364	13	TGCTTAGCTTTCCCTCCTTA	GAGCATTTCTGTTGTTCCTA	45	67	67	-	-	1	1
gs001389	pm2301	13	CATGAACCTGCTCACGACAA	GCCTTACTTTAATGCTGACC	51	100	100	100,>200	-	1	1
gs001492	pm0541	13	AAATGAATGTAAATAGCACT	ATTTAGTTTACAGGGAGAAT	41	72	72	-	74	1	1
gs001367	pm0441	14	GTTTTAAGTTTTGATTTGGG	CATTCCACTCTTACATTTCT.	41	77	77	>200	180	3	4
gs001564	pm2307	14	CGTTCCTAAACTCTGAAATC	AATGCTCATTTATTCTCAAG	42	55	55	>200.	>200	1	1
gs001576	pm2019	14	ATCACAATTACCTTTAGTTG	ACGATAACTTTATTGGAGAT	39	69	69	-	-	1	1
gs001339	pm2220	15	TCCCCATCCTCAGTTGAAGT	TGAGAACAAAAGGAACCAGT	47	70	70	80	150	1	1
gs000980	pm0985	16	TTGGAATGGAACCCTTGCTA	ACTTATGCTGCCTGAAATGG	48	79	79	66	70	2	2
gs001242	pm1127	16	CCCTTGTTTTTACATGTTCA	TATTAAATTCTCCCATTCAT	44	105	105	103	102	2	2
gs001516	pm2543	16	ACAGTGCTAAAATCAAAGGTG	TCTGACAACTCAAGGTGCAAT	45	70	70	>200	-	1	1
gs001566	pm0913	16	TTTGTGTCGGACTATGTAAT	TCACTTTTAATGGGAACCAG	41	53	53	>200	>200	1	1
gs000806	pm1157	17	CTCTCCATGTTCTCTACAAG	TAGAAGGAGAATCTGTGGTT	47	77	77	140	>200	2	3
gs001015	pm2369	17	ATATTCACCTTCCCATCCAT	TCAAATACGTCCTCTCAAGC	50	80	80	>200	>200	1	1
gs001156	pm0202	17	CAGAAATTAAGTGCAGCAAT	TCGTATCTGCATCTTTAAGT	45	103	100	>200	>200	2	2
gs001173	pm2117	17	AAAATCTTGTGGTTATTTCC	GTGATTCTACTGTACATTGC	41	118	118	145	200	1	1
gs001301	pm1878	17	TAAATTTGTGGAAATCTCTTGGA	ACACATTTGGGTTTGCTTTAAC	47	100	100	95	87	1	1
gs001316	pm0514	17	TGTGACAGCAGCAGCTTCAT	TCGTACATTTTAATTCCACC	45	128	128	-	1	1	1
gs001356	pm0538	17	CATCTCACAGACAAGGAAAC	ACCTAAGAGTCCAGAGAAAC	48	90	90	69	>200	1	1
gs001495	pm2212	17	TGACTGCAATAAGGAGTTGT	GAACATACCACGTTTATTTCT	46	90	90	180	>200	1	1
gs001522	pm0642	17	GTCTTCAGCAGATTTCAGGT	ACTTTCTTCTTGAGGACACA	45	68	68	160	-	1	1
gs001078	pm1815	19	TGTGTTCTCCAGCTTTGTAG	GTTACATTGCCTTGGTACAG	48	65	65	>200	>200	1	1
gs001417	pm0289	19	GGATCAGACCAACAGTGCTG	GCAAGGTATAAAACAGATTA	46	50	50	-	-	1	1
gs001467	pm1688	19	GAAGCCCACCCTGCACCTCA	GGAGAGTATTGGGGAACGGT	54	93	93	>200	>200	2	2
gs001069	pm1879	20	GCCATGCTTGTAAAGTGATGT	TTAAGAAGCCATTAGCTAGGATA	48	140	140	-	-	1	1
gs001088	pm1146	20	GCCCTTAGGATTCACTGCTC	ACCACCCAAGGTCTTTCAGG	52	66	66	180	>200	1	1
gs001089	pm0112	20	TGCTGGATGACTTCTACACG	TCCCTATCATGGCTGCTGTT	49	59	59	59,115	59	1	1
gs001128	pm0332	20	CTGCTCGGCTAGTCTGACTC	CAAATGGTCTAAGAGGACAT	49	135	135	153	160	1	1
gs001132	pm0847	20	TCTGAATGATGATGGAAACA	ATCCTAGTCCCAACCCAGTA	48	109	109	-	-	1	1
gs001158	pm1774	20	GGAGCCACATGGATTGATTG	AAATGTACCCCTGGCACCTC	52	124	124	>200	>200	1	1
gs001210	pm1235	20	AGCCATCTGGTTATGTCTTA	GGAGCAGAATGAAACTTCAC	44	90	90	>200	>200	1	1
gs001377	pm1701	20	TCCATGGTGTTAGAAGCCAG	CCACATCTCCAACAGGGAGT	54	142	142	>200	74	1	1
gs001395	pm2101	21	GTCAGCTCAATGCTACACAG	TTTATAGTGCAACACAGAGT	45	130	130	180	>200	1	2
gs001427	pm0648	21	CTTCTGCTATAAAAGTAGAG	ACAATTGGTTCACTAAATGA	39	58	58	145	>200	1	1
gs000978	pm0912	22	GGTGTAGTGTAACCATTTAG	AGTTGCACCCATCTCCTGTC	46	124	124	>200	>200	1	1
gs001444	pm0911	22	GGTCTTGTTCTCCCATCTGT	AGAAAGCCCCAAAGTAGTCC	48	65,80	65	100	125	1	2
gs001473	pm2231	22	TGAGCTGCACTTACCTGTGAGAG	AAGCAGGTTGAGTTGGGTTTTCT	50	94	94	67	135	2	4
gs001479	pm2328	22	TACAGCCCTCCCAGCTAAAC	TTTATTCTGCATCCACTACAA	46	65	65	190	>200	1	1

Fig. 8

EP 0 679 716 A1

gi000999	pm1759	X	CTGCCATAGTTACCTGGATT	TCACCCACCACTATTTAGCA	47	103	103	-	-	1	1
gi001149	pm2180	X	GGAGGGGAGATATAGATTGT	AAAAAAATCCAGAAGACTGA	46	70	70	135	150	1	1
gi001161	pm0608	X	TTCTATAAGTGTGACCAGTT	GGAGGATTTGAGATACACAT	40	85	85	>200	77	1	1
gi001406	pm1294	X	TAATGCCAGTGAATGTTGCGTAA	GTAAAGGTTTATCCTTGCATCAGA	47	82	82	>200	80	2	3
gi001160	pm2289	1,18	ATCCTGCTGAAATACATCTG	GGGGAGAGACATCACATGAC	46	70	70	68	130	1	1
gi001436	pm0113	1,2,12,13,Y	GATCCGATGGGAGTGTAAAT	AATACAAAGCTAAACCACAA	44	69	69	170	-	1	1
gi001404	pm2272	1,2,3,5,8,12,14,17,X	TTGGAATTGACATTCTCTAT	TTTATTGTAACAAAGCAACT	43	130	130	150	132	2	2
gi000803	pm0314	1,2,6,X	TATCAAGCTGAAAATGTCAC	TTACTGAATCCAGCCAACCA	45	93	93	110	-	1	3
gi000140	pm1481	1,3,4,5,8,16	TCCAAATGAAGAAGGTGTTA	AGTTGACAGCCAGGTGAATG	49	96	96	100	100	1	3
gi001354	pm1561	2,20,21,22	GTCTGTCAAGCCAAGATTCA	TTTTTATTGTTGCTCCAAGT	48	110	110	170	150	1	1
gi000336	pm2795	2,4,5,10,12,15,17,20,22,Y	GACCTGTGACATTCTGGACT	TTATATGGTTGTTACACTCG	43	61	61	-	-	2	6
gi001077	pm0943	2,5,14,C	GCCTTGTTATTTCACCACTC	ATCTCCCTTTGCTCCAGTTA	46	82	82	>200	82	1	1
gi001192	pm1853	2,8,12	TCTGAGGACATTCCAAGACAG	CAGTCAAAACCAACACGGTAT	49	95	95	83	160	1	1
gi000213	pm1778	2,9,13,17,X	TGCAATAAAGGGAAAGACCA	CCGTTGTAGGTGATGAAATG	49	78	80	>200	>200	1	2
gi000819	pm0885	20,X	GTCATTTGTATGCAATTTCC	ACATTTTTATTTTTTCAACG	37	45	45	-	-	1	2
gi001109	pm0457	3,10,15	CATGTACTCAGAGGCACTTC	GCAACTACAAATCCCAAACT	50	133	133	>200	150	2	2
gi000071	pm2651	3,4,M	CAGGGACTGGAGCAGGAAAG	GATTTAACCCATTAGGAAGC	50	101	101	101	88	3	3
gi001426	pm2632	3,6	TTAGGAAAATATGGTTAGACAG	ATAGTATTGGGTTGACACAGTA	43	80	80	>200	120	1	1
gi001391	pm1133	3,8	TGGATTTGCTTTACCTTGTT	ACACCCTCAGGAGATGTTAC	47	93	93	95	>200	1	1
gi000077	pm2250	3,9,10,15	GCACTACAAGCCAAATCAGA	CTTCTTAACACCAACAGCAG	50	96	96	>200	125	2	10
gi000605	pm0626	4,6	GGATTTCTATTTGCTGTCAT	GTTTATTGTACGGCATTTAC	44	105	100	>200	>200	2	4
gi001212	pm1234	6,20	GCATTAAACAGGAAACAATA	CTGTCCATGTGGCATAAACC	44	110	110	105	107	1	1
gi001312	pm0606	7,18	AGATGCTAACATTAGGGATA	TTTTAGACATACAGAGGAGT	43	81	81	102	-	1	1
gi001441	pm1253	9,11	CCAGACTACAGGCTGATGGC	CCCTTACCCCAGCAACTCTT	55	75,130	75	>200	>200	1	1
gi001357	pm0115	9,M	ACCAATGTCACTGCTTCTAAAATA	CCCATAATAAGTGAAGAGGTAGTTC	48	125,155	127	125	>200	1	1
gi001261	pm0428	10,15,22	AAGAAATTGTTTACTGGATT	TTATCTGACTTGGAGGAAAT	42	107	107	-	-	1	1
gi001456	pm2420	10,15,22	ACTACCCCTGAGATATTAGTT	TTCATTTATTTGATTAGTTGA	46	100	100	170	-	1	1
gi000250	pm2303	11,M	ATACCACTTCCGCTGTCACG	GAGGAGCGTCTACTGGTCTT	50	72	74	72	>200	3	20
gi000314	pm2643	12,19	GCACCAAGAAGCAGTTCCAG	TTGGGAATGAGAAAATAACT	46	83	83	81	-	1	7
gi000103	pm2773	12,M	GATCTCAGTTCTGCGTTTATT	TACATACAAAGATGCAAACAGT	44	80	80	79	68	1	1
gi001487	pm2725	13,16	ATTCTTGTGTGCTGCTTTCC	GTCTCTCTTCTGATGGCTGA	46	62	60	135	180	1	1
gi000976	pm2780	14,18	AACCTGTTTTACCGCATCTT	AGGTTATTTGTCCACCAGAA	48	87	87	>200	>200	1	1
gi001435	pm1683	17,20,C	TGTTGGTTCACCATTGAGAC	AGAACAACACATCAAAGATGC	46	90	90	>200	90	1	1
gi001393	pm1748	17,22,Y	GAATGTCATCCAAGACGTAG	CTAGTTATATCCTGGCTCTG	44	81	81	>200	200	1	1
gi000056	pm0964	17,C	TTTATCCCAGCAAGCACAAC	TCTTCCTCTTCACTCTCCTC	49	120	120	>200	170	4	11
gi001369	pm2217	17,C	ACTTAAAGTAGCTTTGTACG	TGCCTCCTGGTCTGATAATA	43	95	95	>200	95	1	1
gi001440	pm1213	18,C	CCCCAGTTAAAGATTATTGT	AGTGACGATGGAAGGATGTA	44	92	92	-	92	1	1
gi001217	pm1118	19,20	TGCAGAGTGATTTTCCAGAG	CGTAGGTCATTCTTTTCAGC	46	75	72	160	65	1	1
gi001009	pm2828	19,22	ATCCCTCTGTCTATTCACAC	GCTCGTTTAACTCACTTCAC	46	110	110	130	170	2	2
gi001172	pm0887	19,22	GCCTGCATCTGTGTTGACTT	AACCTCTGGGAACAAATCAT	48	91	89	160	86	1	1

Fig. 9

EP 0 679 716 A1

gs001057	pm2049	C	AGGACACAAACACCAGCTAT	TTTTCTGATTATGACATGAC	45	75	75	101	75	1	1
gs000473	pm1753	M	ATCTCTTTGTAGCCATCCTG	GTTAAAGTGCTGATGCCATT	42	84,100	64	64	>200	1	1
gs001096	pm2236	M	GTAGAGCTGCATTGACTACC	ACAGACAAGGAATAATCATA	42	108,96	110	110	112	1	1
gs001166	pm0506	M	GTCCCACAGTCCAGCCTAAC	GCCACATATTAGAATCCATC	46	74	74	74	>200	1	1
gs001454	pm2354	M	TGTCTTTGTGGACTCTGCCT	TTTAACAGTCAATAAATACATGTT	44	110	110	110	106	1	1
gs000329	pm2492	M_1C	GCTAGAAAGAAGGGCACTCA	CTTAACTCGATAGCCAGGTC	46	75	75	75	75	1	1
gs000253	pm2786	M_1C	CACAAAACAGCAAACTTCAG	ATGGTTATTTTATCAGATTG	41	83	83	82	83	2	3
gs000285	pm1704	M_1C	TCCACCCAGAGAAGCACACT	AATTCATAGGGAATAGGTTC	48	75,130	75	75	75	1	23
gs000302	pm2318	M_1C	TCGAGAAGGACAAAATCACC	GAACAGGGTTAGTCCATTCG	48	58	58	58	58	1	1
gs000543	pm1689	M_1C	CATGAGGCTACGGAAACAGG	AGGAGTCCGTGGGTCTTGAG	51	84	84	84	84	4	18
gs000675	pm1442	M_1C	AAAGCATCTTGAGAGGAACA	GGAGGACTCGCTTGGTCTTA	49	110,>200	110	110	110	1	9
gs000732	pm1452	M_1C	GCAGCAGATACCTTTACACC	TGGTTCATTTCAGTTCCTTC	51	102	105	102	102	3	13
gs000995	pm0268	M_1C	GAAGCTCTTGTGAGGAAAGT	CAGACCCCATCTTTTATACC	47	79	79	79	79	3	4
gs001016	pm2783	M_1C	ACGATATTTATAGTGATGTG	TCAAAACTTTAATATATGCT	40	93	93	91	92	1	1
gs001053	pm1144	M_1C	AGATGAGTGTGGGTCAGAGA	CCATTCCTGTCATTCCAGTT	52	135	140	135	135	1	1
gs001127	pm2290	M_1C	ACTGGTGATGGAAGGTTACA	CCACACAGTGAGCACCGTCT	47	55	55	55	55	1	1
gs001167	pm1626	M_1C	GAGAGCCCTTGCATCCTTTA	CTTCCCTTTGGTCTTTCTGT	49	100	100	100	100	1	1
gs001216	pm2109	M_1C	TAGTCAGAGATTCAGTAAGT	ACATGTATTTTGATAGTCTT	42	110	110	110	110	1	2
gs001253	pm1240	M_1C	AACTGGTTCCATCAAGACTG	AGTGAATAAAGTCTCCACTCC	48	120	120	120	120	1	1
gs001281	pm1131	M_1C	ACTTAAAAACCCACCAGCAT	ACAACAGCAGTCAAATAGAA	47	97	97	97	97	1	1
gs001375	pm0952	M_1C	AAGAGGAGTTTCCCTGCTCA	ATCATGGCAGATGGCAAGGA	51	89	89	89	150	1	1
gs001396	pm2216	M_1C	ATCTGCATGACCTATAATCT	CGTTCTCTTTTATTTGACAT	45	108	108	108	108	1	1
gs001411	pm0958	M_1C	ATGGGTTTATCAGGGGTTTC	GAGACCAAAGGCACTTCTTA	47	80	77	80	80	1	1
gs001460	pm2626	M_1C	ACATTGAATGGGGATGAGGT	GGACATTTCTAGCCCACAGC	51	75,55	75	75	75	1	2
gs001482	pm1210	M_1C	TTGTTGACATTCCTTTTAGAA	CAGTGCCTCTGTACTGAGACA	48	85	85	85	85	1	2
gs001493	pm0109	M_1C	GCCCACAGAGACATCATCCT	TCTTAGTAGGTGCTCTGGTG	51	98	98	98	98	1	1
gs000188	pm2042	No product	CAACCAGTTAGCGTGAAAGT	GAAATAATCCTTGTCATCTA	45	87	87	-	-	1	1
gs000850	pm0304	No product	CTTTGGGATATTTTCTTCAT	CCCTCGGGTACTTTTCTATG	43	60	60	62	-	1	2
gs000983	pm0888	No product	AGCCAGCCTCTTTGTATGTG	CTGGATTTGATTTTCATTAG	44	87	87	-	112	1	1
gs001254	pm1673	No product	TGTGGTATGAAAATATCTGA	TTATGAATGAAGACAACACT	43	98	98	183	>200	1	1
gs001365	pm2908	No product	CAGTAGTGTGCTTTGAAATG	TTTATGTGAAATGTGGTTGT	41	63	63	-	150	3	3
gs001373	pm0361	No product	TACAGCCGCTTCTAAAAGTC	TTTGAGCATCAAGGAAATCT	46	82	82	-	>200	1	1
gs001558	pm0849	No product	TACATTCTTCAGACTCATCG	TTTTCAAAAACTTTATTCTT	40	86	86	>200	100	1	1
gs001574	pm1284	No product	ATCAGAGCTCAGTTCCTGTAG	ATTTTGCCTCTTGCATGGTC	44	57	57	67	67	2	2
gs001622	pm1606	No product	GATCTTGAGCCTTAACTGGA	TTTGCAGTTCAGCTTTATTC	45	54	54	-	-	1	1
gs001640	pm0852	No product	GATCTCTGTTCCTTTTCACA	TTTATAACAAGACACCATAC	36	45	45	-	-	1	1

Fig. 10

EP 0 679 716 A1

0179

DNIC

Fig. 11

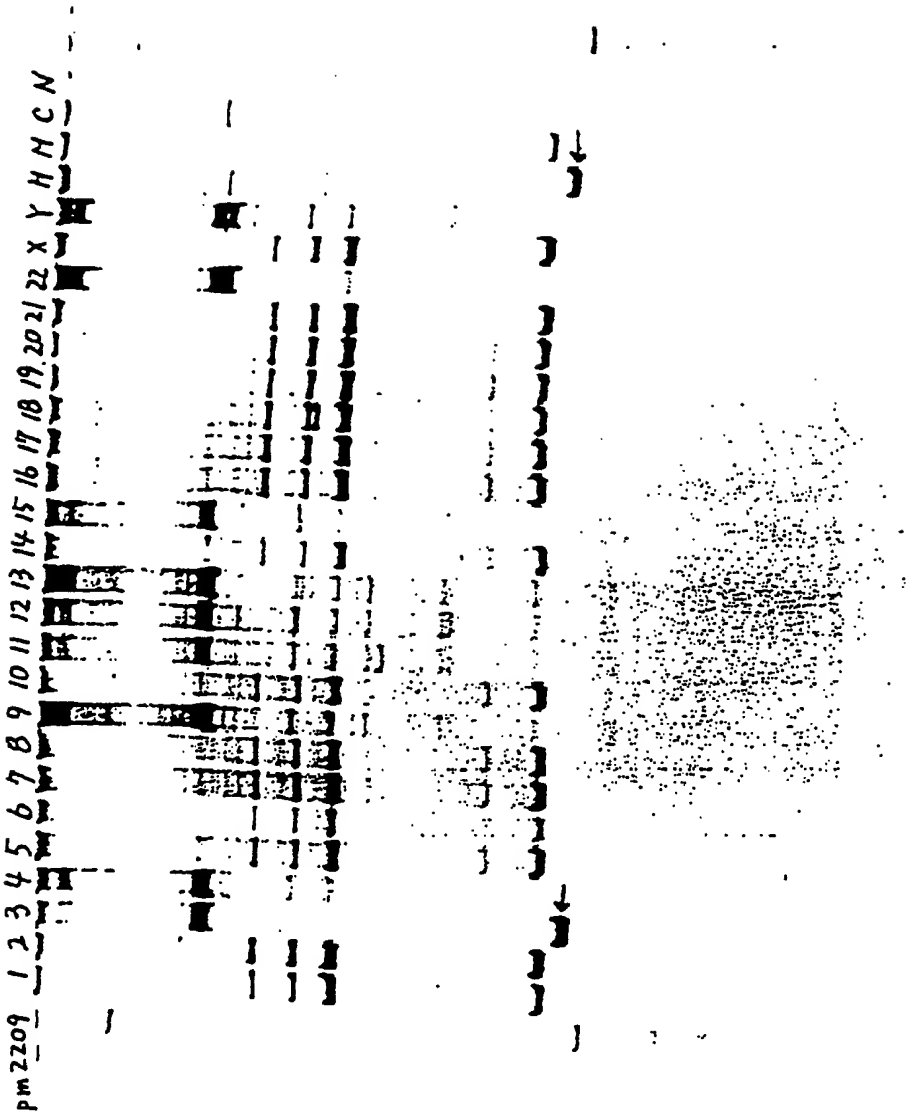


Fig. 12

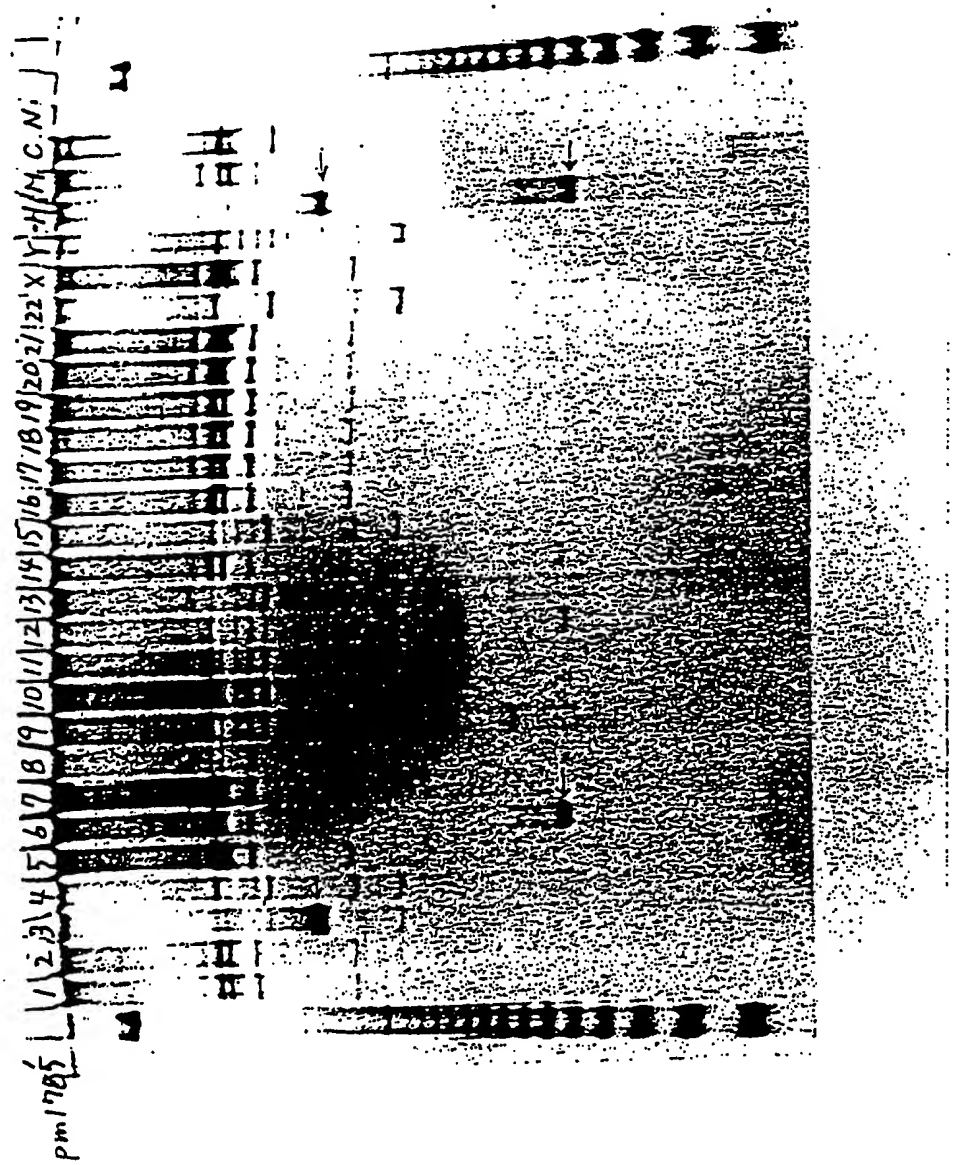


Fig. 13

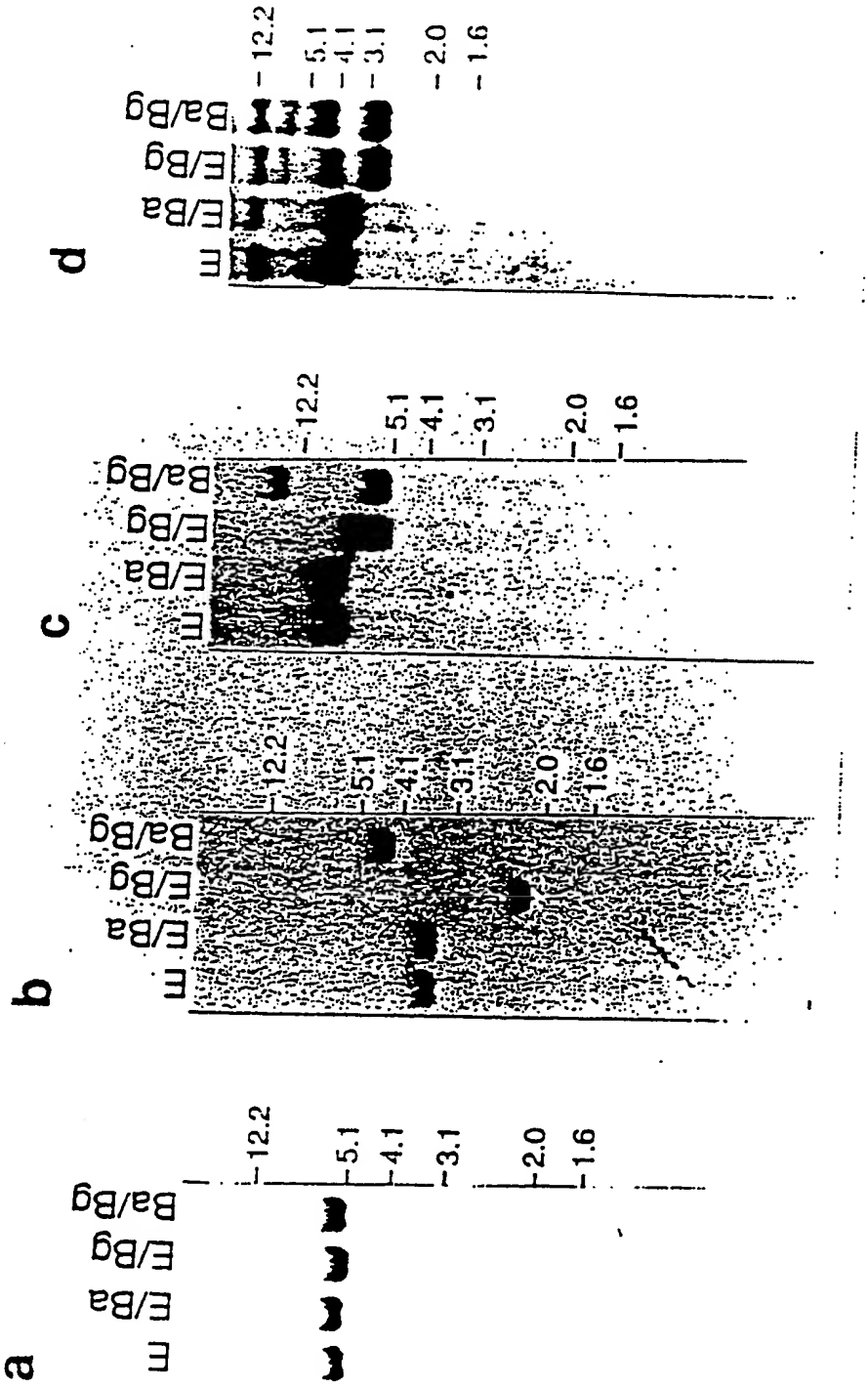


Fig. 14

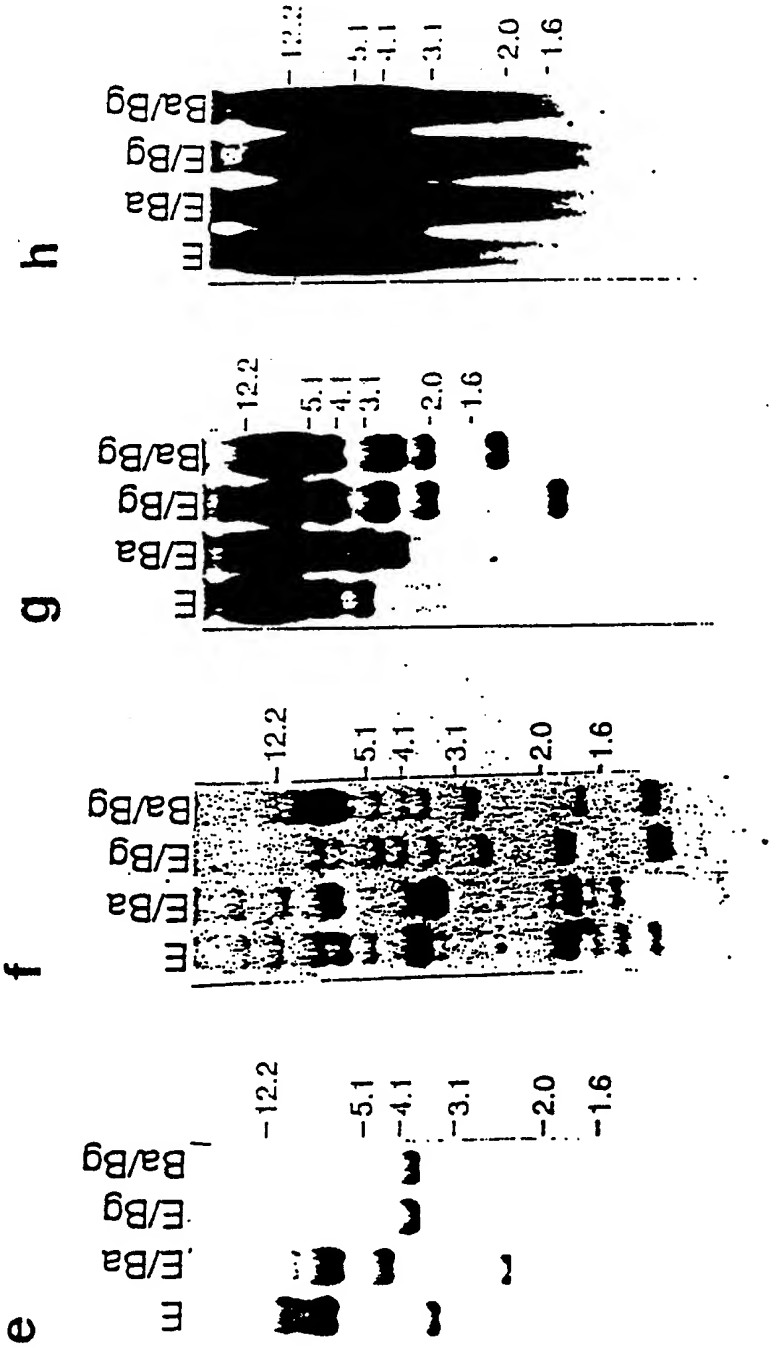


Fig. 15

Hybrid cells used for Southern hybridization

Hybrid cell	Human chromosome No.	Parent cell	Intact chromosome (%)	Translocated chromosome (%)
A9(neo-1)-4	1	A9	100 (0)	0
A9(neo-2)-1	2	A9	93 (8)	0
GM10253	3	CHO	100 (0)	0
GM10115	4	CHO	100 (0)	0
A9(neo-5)-4	5	A9	40 (0)	90
A9(neo-6)-3	6	A9	100 (60)	0
A9(neo-7)-2	7	A9	100 (89)	0
A9(neo-8)-1	8	A9	91 (82)	0
GM10611	9	CHO	79 (5)	11
A9(neo-10)-3	10	A9	94 (6)	75
A9(neo-11)-1	11	A9	24 (0)	76
GM10927A *	11	CHO	96 (21)	4
A9(neo-12)-4	12	A9	0 (0)	100
GM10868 *	12	CHO	82 (6)	0
GM10898	13	CHO	82 (0)	10
GM10479	14	3T6	76 (29)	0
A9(neo-15)-2	15	A9	9 (0)	78
GM11418 *	15	CHO	62 (0)	100
GM10567	16	A9	69 (0)	0
GM10498	17	LTMK	80 (10)	0
A9(neo-18)-5	18	A9	100 (66)	0
A9(neo-19)-1	19	A9	92 (23)	8
A9(neo-20)-3	20	A9	81 (5)	17
GM08854	21	A9	81 (24)	0
GM10027	22	CHO	93 (0)	100
GM10324	X	A9	81 (10)	0
GM06317	Y	CHW1103	91 (0)	9

Fig. 16

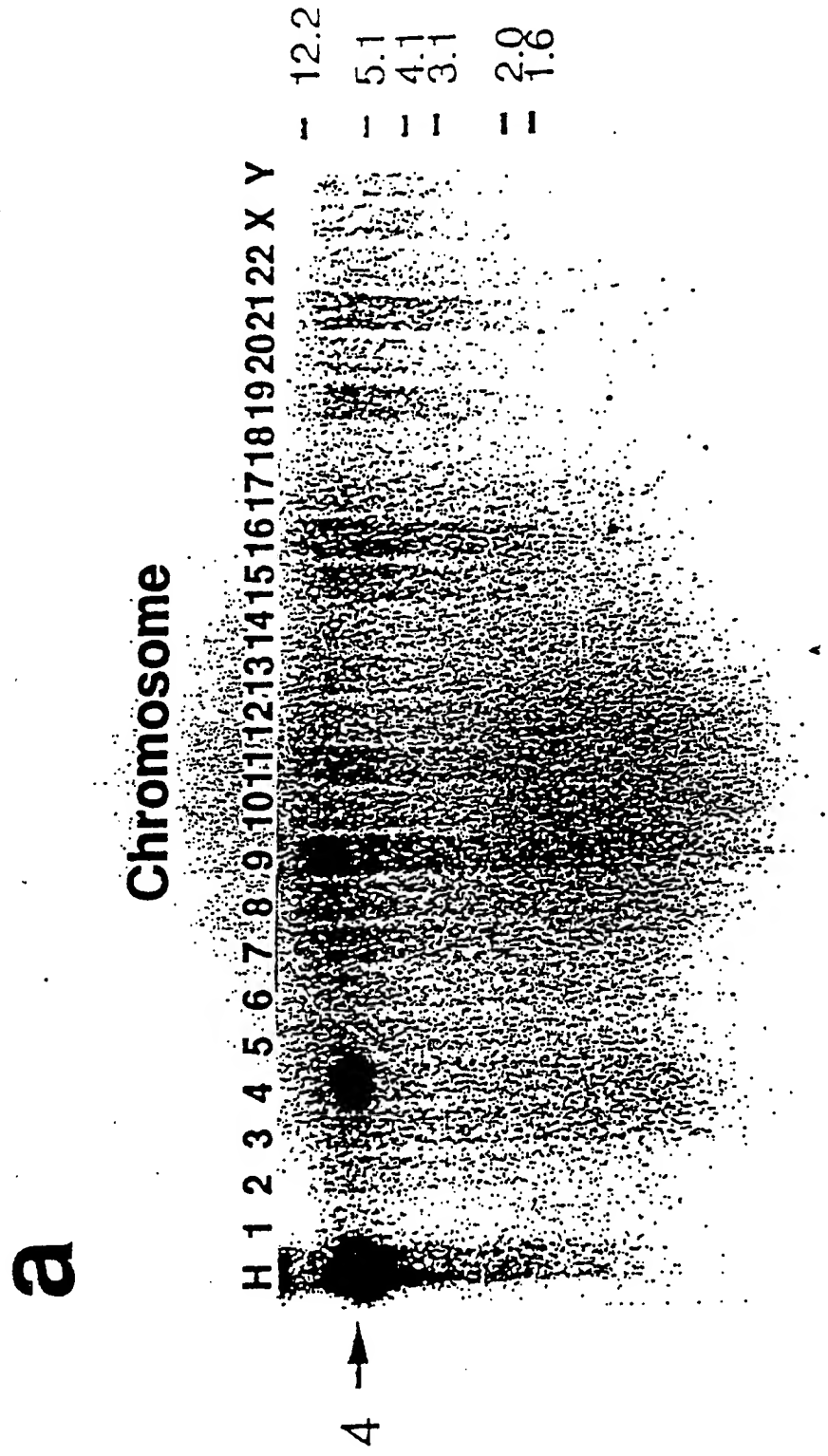

a
Chromosome
H 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
12.2
5.1
4.1
3.1
2.0
1.6
4

Fig. 17

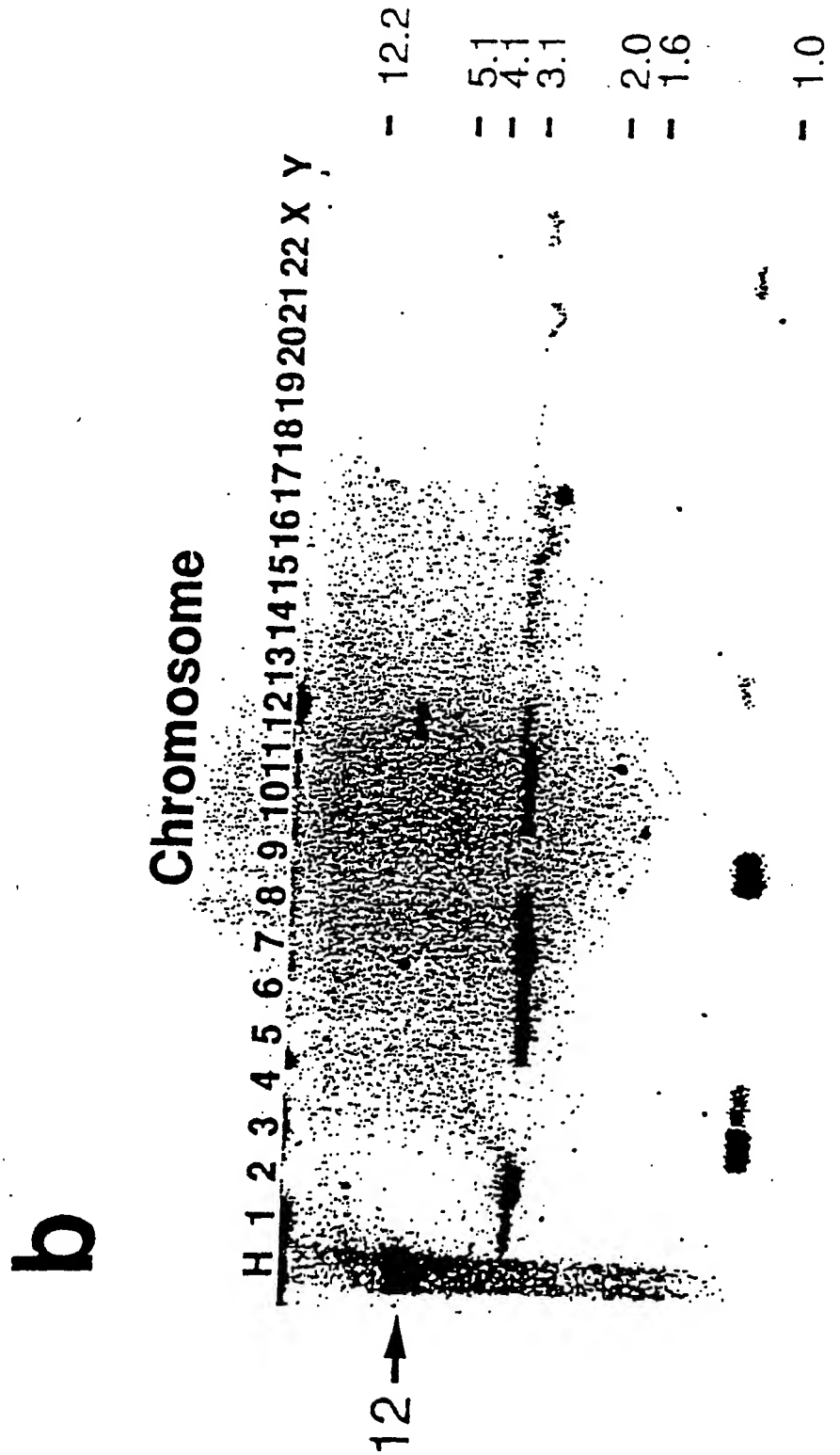

b
Chromosome
H 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
12.2
5.1
4.1
3.1
2.0
1.6
1.0
12

Fig. 18

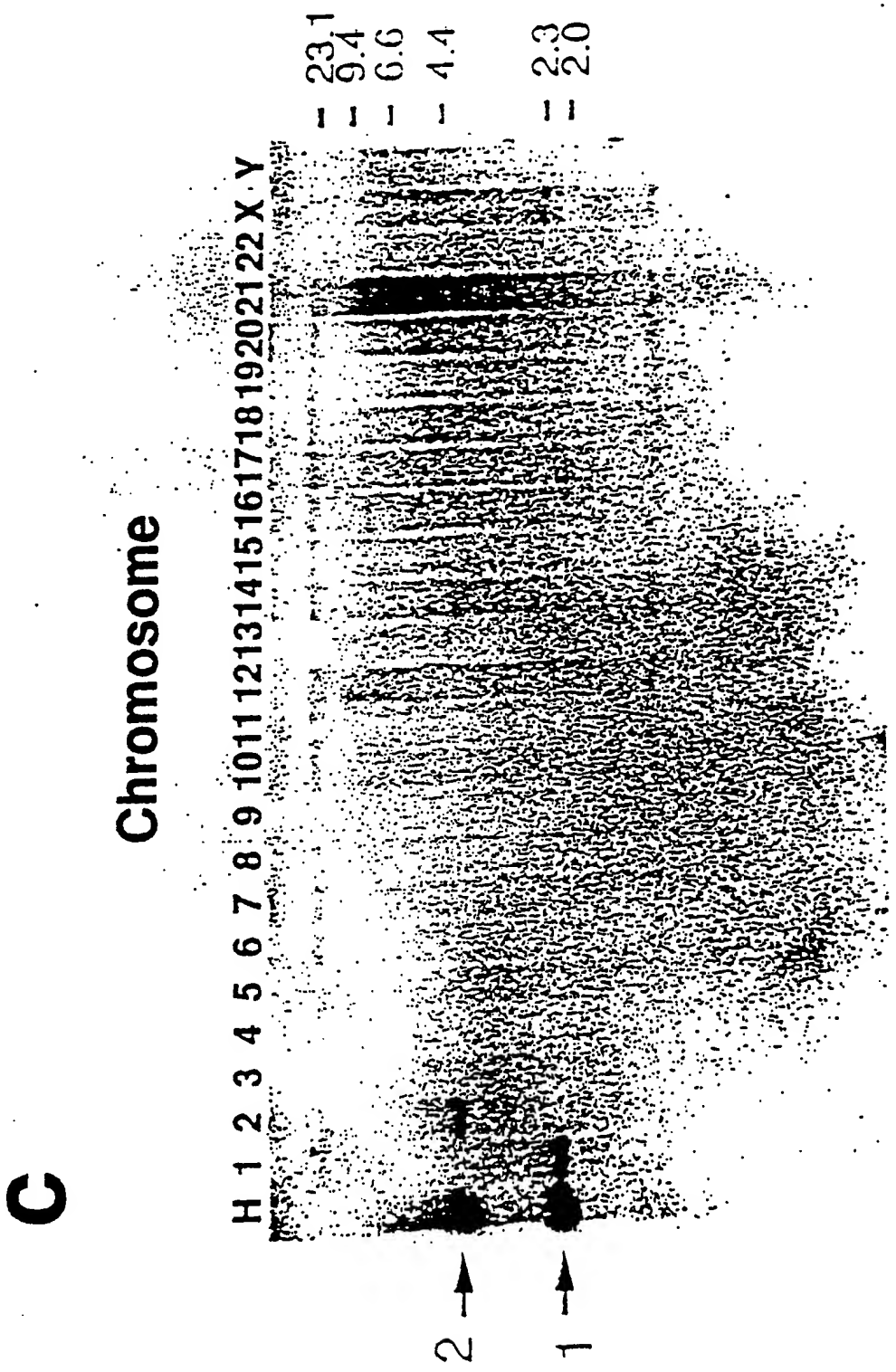

C
Chromosome
H 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
23.1
9.4
6.6
4.4
2.3
2.0
2
1

Fig. 19

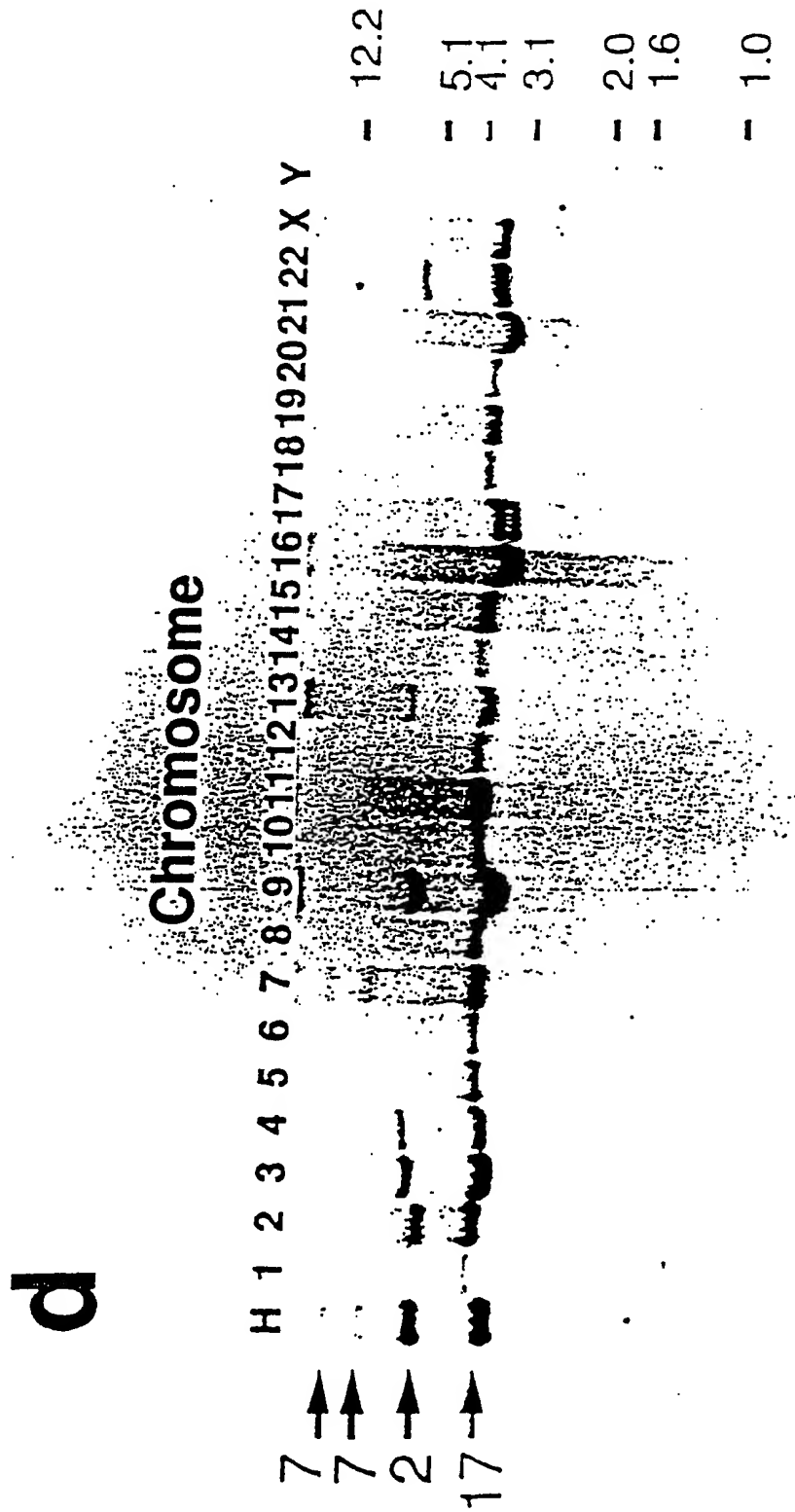

d
H 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosome
12.2
5.1
4.1
3.1
2.0
1.6
1.0
7
7
2
17

Fig. 20

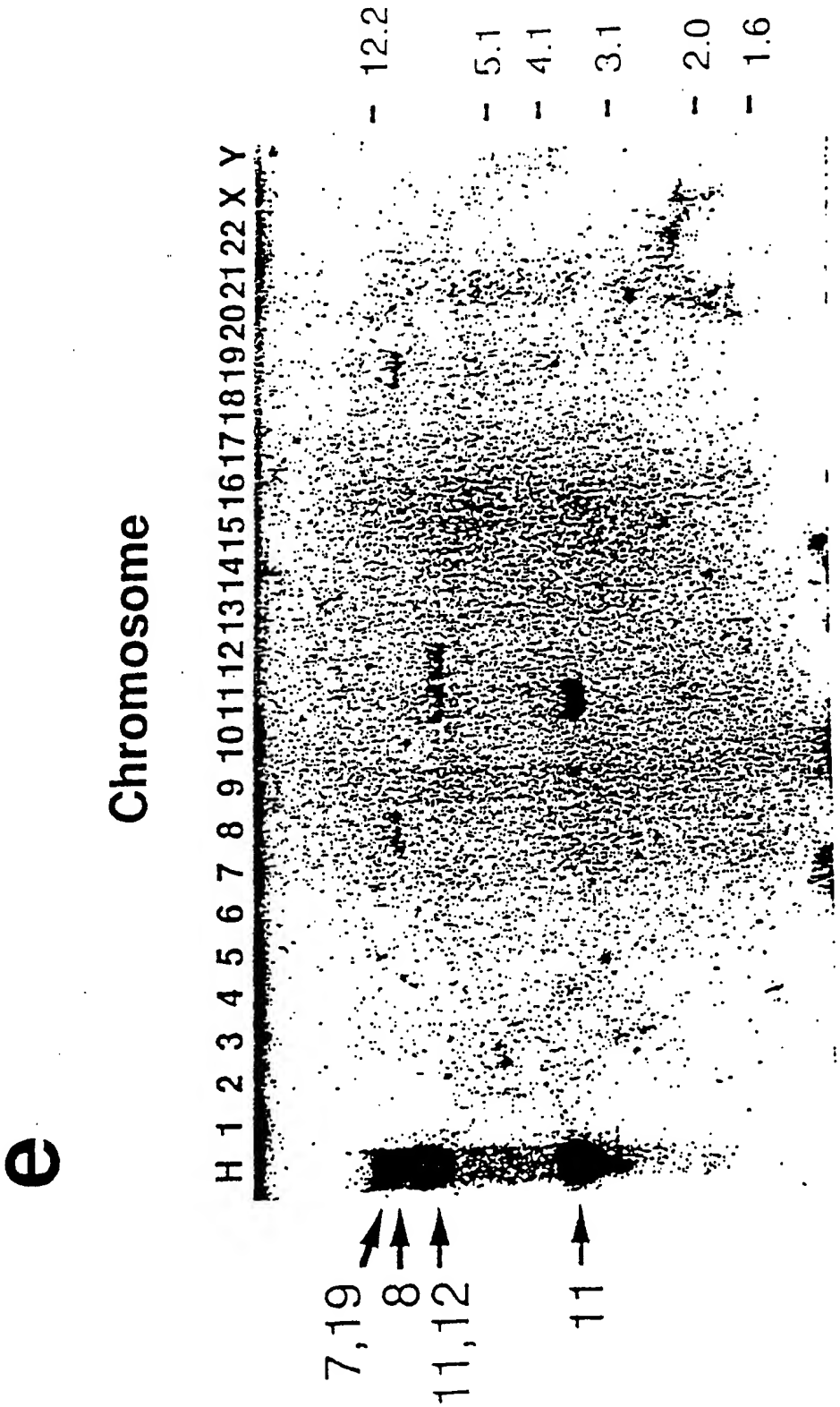

e
Chromosome
H 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
12.2
5.1
4.1
3.1
2.0
1.6
7,19
8
11,12
11

Fig. 21

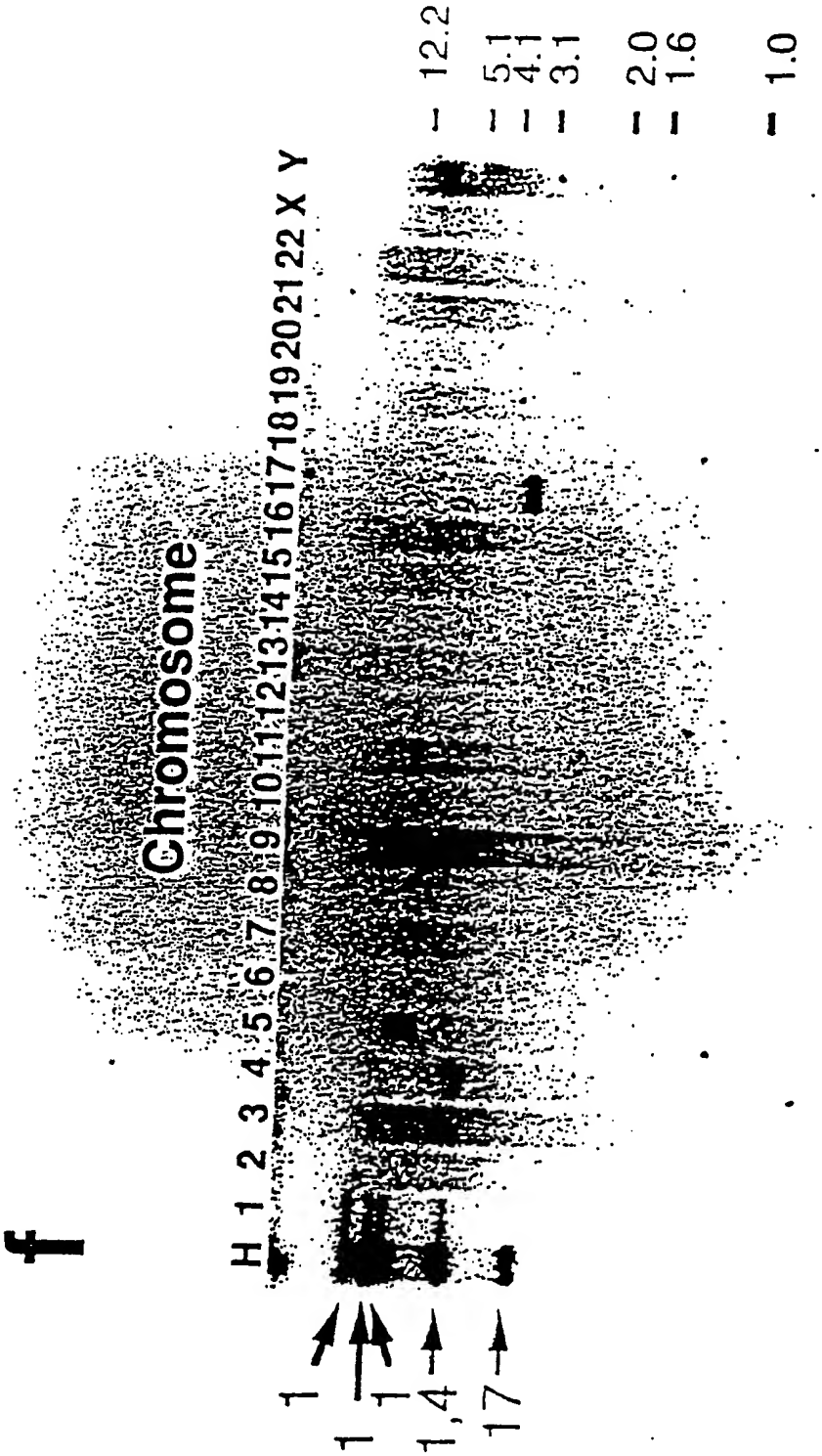

f
H 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosome
12.2
5.1
4.1
3.1
2.0
1.6
1.0
1
1
1,4
17

Fig. 22

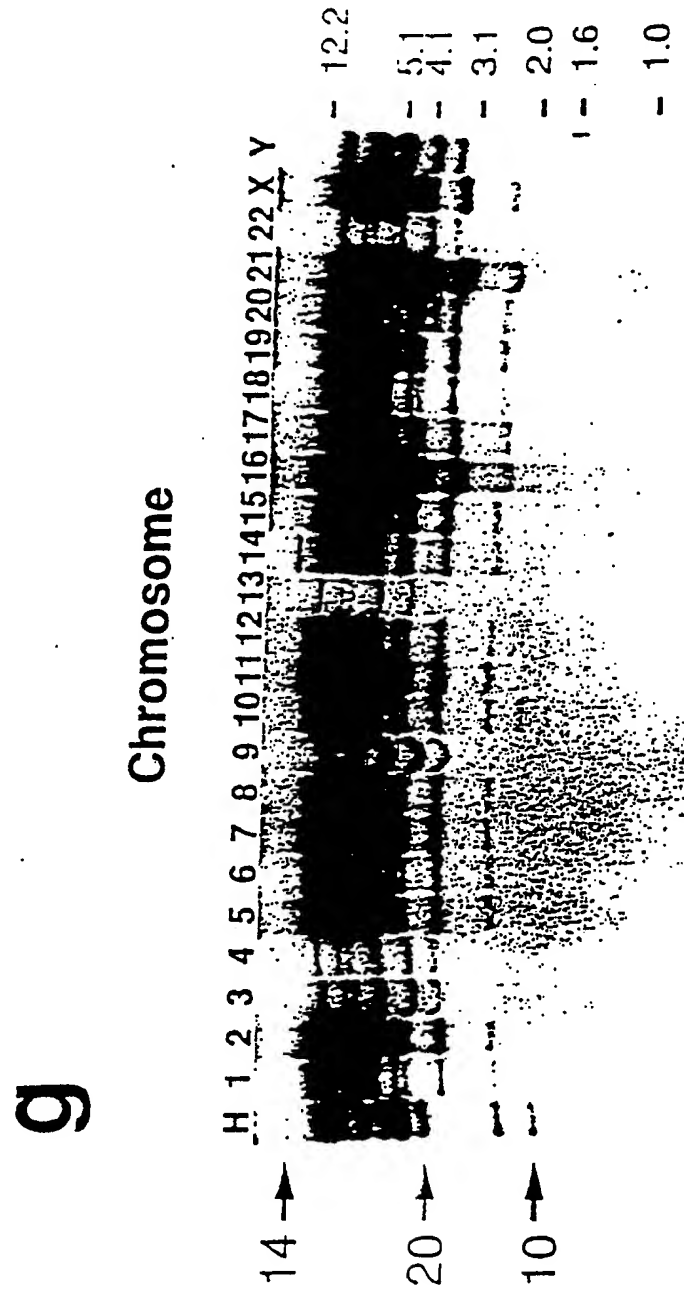

g
Chromosome
H 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
12.2
5.1
4.1
3.1
2.0
1.6
1.0
14
20
10

Fig. 23

Chromosomal mapping of each GS by Southern blot technique

Clone		Sequence length	Numbers of bands detected with human whole chromosomes: E	E/Bam	E/Bgl	Bam/Bgl	Chromosomes assigned	Background: Mouse	Background: Chinese hamster
Single band group:									
c12c11	GS000075	432	1	1	1	1	9	0	0
c12c06	GS000062	540	1	1	1	1	6,15	0	0
c12g01	GS000290	212	1	1	1	1	2	1	1
c13c05	GS000117	359	1	1	1	1	11	0	0
c13c07	GS000120	355	1	1	1	1	2	0	0
c13f10	GS000206	267	1	1	1	1	14	0	0
c13h01	GS000279	183	1	1	1	1	12	0	0
c13h02	GS000322	167	1	1	1	1	6	0	0
d0g02	GS000095	397	1	1	1	1	3	0	0
d0h07	GS000184	313	1	1	1	1	11	1	1
d1b10	GS000348	153	1	1	1	1	20	0	0
hm01a12	GS000223	246	1	1	1	1	27	0	0
hm01c09	GS000428	157	0	1	1	1	1	0	0
hm01e12	junk	394	1	1	1	1	17	0	0
hm01f05	GS000066	454	1	1	1	1	19,22	0	0
hm01f10	GS000299	173	0	1	1	1	10	0	0
hm01g09	GS000053	477	1	1	1	1	6	0	0
hm01h07	GS000115	368	1	1	1	1	12	0	0
hm02a02	GS000130	344	1	1	1	1	4	0	0
hm02a04	GS000329	164	1	1	0	0	10	0	0
hm02e01	GS000203	271	1	1	1	1	16	0	0
hm02e01	GS000016	590	1	1	1	1	20	0	0
hm02e02	GS000342	156	0	1	1	1	14	0	0
hm02e05	GS000401	223	1	1	0	0	n.d.	0	0
hm02g02	GS000191	278	1	1	1	1	17	0	0
hm05a05	GS000251	219	1	1	1	1	6	2	0
hm05a10	junk	392	1	1	1	1	1	1	1
hm05e10	GS000009	606	1	1	1	1	1	0	0
kmd01	junk	169	1	1	1	0	n.d.	0	0
s105	GS000001	703	1	1	1	1	5	0	0
s110	GS000057	471	1	1	1	1	8	0	0
s11d11	GS000307	#175	0	0	0	1	7	0	0
s11h01	GS000269	204	1	1	1	1	8	0	0
s147	GS000060	461	1	1	1	0	2	0	0
s14c06	junk	639	1	1	1	1	1	0	0
s14g02	GS000152	322	1	1	1	1	4	0	0
s14h12	GS000271	193	1	1	1	1	4	1	1
s150	GS000143	330	1	1	1	1	17	0	0
s156	GS000002	306	1	1	1	1	2	1	1
s15b11	GS000250	221	1	1	1	1	14	0	0
s179	GS000275	196	1	1	1	1	n.d.	0	0
s246	GS000234	241	1	1	1	1	9	0	0
s247	GS000347	153	1	1	1	1	1	0	0
s270	junk	185	1	1	1	1	19	0	0

Fig. 24

Clone		Sequence length	Numbers of bands detected with human whole chromosomes: E	E/Bα	E/Bβ	Bα/Bβ	Chromosomes assigned	Background: Mouse	Chinese hamster
s306	GS000256	205	1	1	0	1	X	0	0
s309	GS000171	305	1	1	0	1	1	0	0
s342	GS000328	165	1	1	1	1	4	3	2
s331	GS000255	207	1	1	0	1	6,15	1	1
s334	GS000165	312	1	1	1	1	1	0	0
s337	GS000278	195	1	1	1	1	17	0	0
s339	GS000295	180	1	1	1	1	n.d.	0	1
s443	GS000330	251	1	1	1	1	n.d.	0	0
s470	junk	251	1	1	1	1	17	0	0
s474	GS000192	273	1	1	1	1	5	0	0
s503	junk	312	1	1	1	1	12	0	0
s507	junk	600	1	1	1	1	1	2	1
s517	GS000334	161	1	1	1	1	14	1	1
s632	junk	587	1	1	1	1	2	0	0
s633	GS000166	311	1	1	1	1	22	2	1
s650	GS000041	644	1	1	1	1	12	1	1
tw1-04	GS000025	537	1	1	1	1	3,7	0	0
tw1-19	GS000218	255	1	1	1	1	17	0	0
tw1-32	junk	250	1	1	1	1	5	0	0
tw1-37	GS000237	235	1	1	1	1	22	0	0
tw1-42	junk	391	1	1	1	1	8	1	1
tw1-43	GS000098	178	1	1	1	1	14	0	0
tw1-96	GS000133	339	1	1	1	1	11	0	0
Two band group:									
c12f12	GS000195	277	1	2	2	2	1,	1	1
c13d02	GS000042	503	2	2	1	1	2,	0	0
hm01a06	GS000129	344	2	2	2	2	11,18	3	5
hm01a07	GS000207	269	2	2	2	2	7,	0	0
hm01d05	GS000232	243	2	2	2	1	2,	0	0
hm01e01	GS000181	292	2	2	2	2	1,2	0	0
hm02a08	GS000435	302	2	2	2	2	3,	1	1
hm02e04	GS000221	253	2	2	2	2	3,	0	0
hm02e05	GS000146	332	2	2	2	2	17,19,22	0	0
hm05f07	GS000043	503	1	1	2	1	3,	0	0
s11d06	GS000268	205	2	2	2	2	11,12	0	0
s11g12	GS000337	255	2	2	2	2	6,	0	0
s124	GS000083	404	2	2	2	2	9,	1	1
s144	GS000132	342	1	2	2	2	1,7	0	0
s14f03	GS000239	243	1	2	2	2	2,	3	2
s15e02	junk	439	2	2	1	2	6,	0	0
s16b09	junk	420	1	1	1	2	10,14	0	0
s17c09	GS000248	223	2	2	2	2	14,	0	0
s231	junk	284	2	2	2	2	11,	0	0
s254	GS000124	353	2	2	2	2	1,	3	1
s255	GS000235	239	2	2	2	2	11,	0	0
s272	junk	195	2	2	2	2	10,16	1	1

Fig. 25

Clone		Sequence length	Numbers of bands detected with human whole chromosomes: E	E/Bs	E/Bg	Bs/Bg	Chromosomes assigned	Background: Mouse	Chinese hamster
s311	GS000092	333	1	1	2	2	16.	1	1
s313	junk	132	2	2	1	0	20.	0	0
s317	GS000100	339	0	0	1	2	14,14	1	1
s336	GS000134	337	2	2	2	2	12,14	0	0
s338	GS000139	283	2	2	2	1	22,X	0	0
s339	GS000293	137	2	1	1	2	17.	0	0
s394	GS000068	449	2	1	2	2	13,14	0	0
s396	junk	277	2	2	2	2	17.	0	1
s455	junk	452	1	2	2	1	4.	0	0
s456	GS000236	132	2	2	2	2	8,10	1	2
s465	GS000201	274	1	1	2	2	6,15	0	0
s635	junk	250	1	1	1	2	9,13	0	0
s639	GS000257	205	1	2	2	2	2,X	0	0
s656	GS000025	#590	2	2	0	2	6,11	0	0
tw1-33	junk	352	2	2	2	2	1.	0	0
tw1-39	GS000153	#321	2	2	2	2	17.	0	0
tw1-70	GS000061	441	1	1	2	1	11.	0	0
tw1-80	junk	453	2	2	1	2	9,17	2	2
tw1-87	GS000158	316	2	2	2	2	7.	0	0
Three band group									
d0h06	GS000080	417	3	3	3	1	1.	0	0
hm05b07	junk	336	2	3	3	3	5.	0	0.
hm05g02	GS000209	267	2	2	2	1	3,17,19	1	1
s129	GS000107	378	3	3	3	3	n.d.	1	1
s178	GS000357	146	1	2	2	3	2,	0	0
s17a10	GS000294	131	3	3	3	3	2,13,22	1	1
s308	GS000412	638	2	2	2	3	XX	1	1
s401	GS000224	249	2	3	3	3	6,6,	0	0
s654	GS000045	491	3	3	3	3	1,22,	0	0
tw1-82	GS000208	267	3	3	3	3	13,	4	0
Four band group :									
c12g07	GS000154	320	4	4	2	3	5, 14,	0	0
c13a08	GS000055	508	3	3	4	4	2,7,7,17	1	2
c13c04	GS000106	#376	4	3	3	3	n.d.	0	2
c13e09	GS000302	195	4	2	4	4	2,17,	7	2
s136	GS000160	315	4	4	4	4	4,X,	2	1
s163	GS000004	#618	4	4	4	2	4,4,8,20	3	1
s479	GS000130	293	4	4	2	2	7,8,11,11,12,19	0	0
Group with 5 or more bands									
c12f08	GS000253	217	5	5	5	2	2,7,9,14,	2	0
he01	junk	374	12	12	15	13	1,2,6,	22	20
hd10	junk	361	4	4	4	8	n.d.	12	6
he10	junk	173	6	2	3	3	6,8,9,19,21,	3	3
hm01c05	GS000205	176	9	7	5	5	X	9	8
hm01f04	GS000246	215	8	10	5	5	n.d.	12	12
hm01g02	junk	411	9	6	6	4	10,14,20,	14	6

Fig. 26

Clone		Sequence length	Numbers of bands detected with human whole chromosomes: E	E/Ba	E/Bg	Ba/Bg	Chromosomes assigned	Background: Mouse	Chinese hamster
hm02f09	GS000273	442	8	7	7	5	3,3,6,11,13,14,15,16	0	0
hm05a02	GS000096	373	5	6	4	6	2,3,17,	3	3
hm05a04	GS000236	#239	6	6	6	7	n.d.	8	5
kmb01	junk	350	3	5	5	5	13,	14	7
s11f06	GS000315	170	6	6	6	4	1,2,2,3,4,6,13,15,	0	3
s14f01	GS000407	262	12	11	10	9	1,6,9,13,	6	3
s173	GS000094	397	5	4	6	8	1,1,1,1,4,17	0	0
s255	GS000323	167	10	12	11	14	13,	9	5
s341	junk	494	9	9	8	6	n.d.	15	3
s406	GS000113	364	6	7	5	4	2,7,8,13,20,20	4	1
tw1-46	junk	593	9	10	10	10	1,1,2,2,5,11,X,	3	5
tw1-63	junk	203	8	10	10	12	3,4,	17	11
Bands no detected:									
e13g02	GS000340	157	0	0	0	0	-	-	-
hm01e10	junk	232	0	0	0	0	-	-	-
hm02d11	GS000274	196	0	0	0	0	-	-	-
s323	GS000273	194	0	0	0	0	-	-	-
s359	GS000199	279	0	0	0	0	-	-	-
s511	junk	233	0	0	0	0	-	-	-
s645	GS000012	#734	0	0	0	0	-	-	-
s647	GS000105	360	0	0	0	0	-	-	-
s651	junk	540	0	0	0	0	-	-	-